Network Approaches for Exploring Predicted Proteins of Unknown Function in the Sequenced Genome of Plant Pathogenic Fungi

Janowska-Sejda, Elzbieta Iwona

Awarding institution:
King's College London

NETWORK APPROACHES FOR EXPLORING PREDICTED PROTEINS OF UNKNOWN FUNCTION IN THE SEQUENCED GENOMES OF PLANT PATHOGENIC FUNGI

Elżbieta Iwona Janowska-Sejda

A Thesis Submitted for the Degree of Doctor of Philosophy

Department of Informatics

King's College London

June, 2018

Abstract

Diseases caused by plant pathogenic fungi have a major impact on crop production leading to local or even global food and feed shortages. In addition, secondary metabolites produced by several pathogenic fungi can cause serious health problems in animals or even humans. Therefore, the identification of virulence genes in plant pathogenic fungi is of huge importance since it would help to understand the infection process and aid in the development of control strategies.

With the advent of new sequencing technologies, the number of available whole-genome sequences and predicted proteomes is rapidly increasing for numerous plant pathogenic fungi. However, there are still many proteins with no assigned molecular function. At the same time, high-quality classification of protein families/domains and mutant phenotype information is increasingly available from databases such as PFAM and the Pathogen-Host Interactions database (PHI-base), respectively. The main objective of this work is to explore in depth proteins of unknown function and thereby speculate on their roles in virulence.

In this study, various computational network approaches have been applied to integrate available biological data for selected eukaryotic pathogens. The Markov Cluster Algorithm was implemented to detect plant pathogen-specific and animal pathogen-specific gene clusters. Further, a neighbourhood-based network analysis approach was combined with a domain-domain interaction (DDI) and interologs high confidence network analysis to predict candidate genes for virulence in a globally important cereal-infecting and mycotoxin producing plant pathogenic fungus, namely *Fusarium graminearum*. Collectively, these analyses newly assigned 65 proteins a role in virulence. Most of those predicted proteins are thought to be a part of the Mitogen-activated protein kinase signaling pathways activated in *F. graminearum* during wheat ear infection. One gene, namely FGSG_06444, was identified to be a high-priority candidate for further biological experiments.

Another new computational approach carried out in this work was the application of the domain-association network to the functional prediction of Domains of Unknown Function (DUFs). Here

available phenotypic data for gene mutants curated in the PHI-base was integrated with taxonomic information, as well as topological properties of protein domains. Results from this novel analysis rejected the hypothesis that certain DUFs are linked to the virulence process of fungal plant pathogens. However, several DUFs were assigned a role in core metabolism (essential for life proteins) instead. Furthermore, a taxonomical diversity study of domains and Louvain community clustering identified 35 DUFs to be fungal-specific domains. A novel life-strategy-integration-analysis was developed where biological information from species employing saprophytic, heterotrophic and biotrophic lifestyles can be integrated into the one platform. This was achieved by combining a Protein Bigrams Overlap Network approach with SimMod analysis. Here two *M. oryzae* proteins (MGG_09419 and MGG_03468) were identified as novel effector protein candidates and six additional *F. graminearum* proteins were identified as members of polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) secondary metabolite pathways (FGSG_00036, FGSG_04588, FGSG_05321, FGSG_10464, FGSG_17387 and FGSG_17677). These genes are likely to be involved in virulence and were suggested to Rothamsted experimental biologists to be tested in gene deletion experiments to confirm their function.

Overall, this study has implemented different approaches to investigate and assign molecular and biological functions to unannotated proteins in plant pathogenic fungi. By employing graph theory to integrate and analyse functional domain information from PFAM database and mutant phenotype information from PHI-base, it is possible to identify candidate genes responsible for virulence in fungal pathogens.

Acknowledgements

I would like to thank my Rothamsted Research supervisors, Prof. Kim Hammond-Kosack and Dr Martin Urban for their invaluable support and guidance throughout my PhD training. I am particularly grateful for their patience and encouragement during the course of this project. I also would like to thank my university supervisor Dr Sophia Tsoka for her support and advice all the way through this study. I am also grateful to Dr Mansoor Saqi for his supervision and advice through the first years of my study. I also acknowledge Dr Stephen Powers for his statistical advice.

I am also thankful to the following people for their support, advice and encouragement: Dr Michael Defoin-Platel, Dr Artem Lysenko, Dr Laura Bennett and Dr Matthew Hindle.

I am also grateful to fellow PhD colleague at Kings College London Jonathan Silva for his help in generating modules from SimMod algorithm, used in Chapter 7 of this thesis.

An exceptional note of appreciation goes to my husband Andrzej for his priceless support throughout the whole time of this research, as well as to Mateusz, Natalia and Anna for being wonderful and understanding children.

Finally, this thesis is dedicated to the memory of my mum, Krystyna.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

3did	3D Interacting Domains
aa	Amino acids
AIC	Average Information Content
AIC-MICA	Average Information Content of the Most Informative Common Ancestor
AllPath	All pathogens group consisting of PP, SP, FP and AP
AP	Animal pathogens (fungi attacking animals)
BioGRID	Biological General Repository for Interaction Datasets
BioP	Biological process, an aspect of Gene Ontology
BLAST	Basic Local Alignment Search Tool
BLASTP	Basic Local Alignment Search Tool for Proteins
BROAD	Broad Institute, Boston, USA
CATH	Class Architecture Topology Homologous superfamily
CC(s)	Connected component(s)
CellC	Cellular component, an aspect of Gene Ontology
CHP	Conserved hypothetical protein
CPBON	Combined Protein Bigrams Overlapping Network
CPGR	The Comprehensive Phytopathogen Genomics Resource
Cytoscape	Bioinformatics software tool for visualising networks
DAG	Directed acyclic graph
DDBJ	DNA Data Bank of Japan database
DDI	domain-domain interactions
DIP	Database of Interacting Proteins
DNA	Deoxyribonucleic acid
DON	Deoxynivalenol
DUF(s)	Domain(s) of Unknown Function
EBI	European Bioinformatics Institute
EC	Enzyme Commission
ETI	Effector-triggered immunity
ETS	Effector-triggered susceptibility
FASTA	Text-based format for storage of protein and nucleotide sequences
FC	*Fusarium culmorum*
FEB	Fusarium ear blight (disease caused by *Fusarium* fungi*)*
FG	*Fusarium graminearum*
FG3	*Fusarium graminearum* third genome assembly
FGDB	*Fusarium graminearum* Genome Database
FGRRES	*Fusarium graminearum* Rothamsted Research genome assembly
FP	Fungal pathogens (fungi attacking other fungi)
FPPI	*Fusarium graminearum* protein-protein interaction
FRAC	Fungicide Resistance Action Committee
FunCat	Functional catalogue classification scheme developed at MIPS
FV	*Fusarium venenatum*
GEN-AU	GENome Research in Austria
GO	Gene Ontology

HMM	A hidden Markov model
HMMER	Bio-sequence analysis using profile hidden Markov models
HP	Hypothetical protein
HPPPI	Host pathogen protein-protein interaction
HPRD	The Human Protein Reference Database
IC	Information Content
InParanoid 8	Algorithm that finds orthologous genes and paralogous genes
IntAct	IntAct molecular interaction database
iPfam	A database that catalogues Pfam domain interactions based on known 3D structures that are found in the Protein Data Bank
KBDOCK	Spatial classification of 3D protein domain family interactions
KDP	Kernel Density Plots
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KS	Kolmogorov-Smirnov statistical test
LS	Lineage-specific
MAPK	Mitogen-activated protein kinase
MCL	Markov Cluster Algorithm
MICA	Most Informative Common Ancestor
MINT	Molecular INTeraction database
MIPS	Munich Information Center for Protein Sequences
MO	*Magnaporthe oryzae*
MolF	Molecular function, an aspect of Gene Ontology
MPPI	*Magnaporthe oryzae* protein-protein interaction
mRNA	Messenger RNA
NC	*Neurospora crassa*
NCBI	National Center for Biotechnology Information
NetworkX	Python package for network analysis
NP	Non-pathogenic fungi
nr	NCBI non-redundant protein sequences database
NRPS	Non-ribosomal peptide synthetase
Ondex	Data integration and graph visualisation framework
PAMP	pathogen-associated molecular patterns
PBON	Protein Bigrams Overlap Network
Perl	High-level scripting language
PFAM	The protein families database
PHI-base	Pathogen-Host Interactions database
PIR	Protein Information Resources
PKS	Polyketide synthase
PLEXdb	Plant Expression Database
PON	Protein Overlap Network
PP	Plant pathogens (plant pathogenic fungi)
PPI	Protein-protein interaction
PPIN-1	Predicted plant pathogen immune network
PRRs	Pattern recognition receptors
PSI-BLAST	Position Specific Interactive BLAST
PTI	PAMP-triggered immunity

PubMed	Database of scientific publications
Python	Interpreted scripting language
R	Statistical programming language /or environment
REMI	Restriction enzyme mediated integration
RIP	Repeat-induced point mutation
RNA	Ribonucleic acid
RWR	Random walk with restart algorithm
SCOP	Structural Classification of Proteins
SimMod	Algorithm for detection composite communities (modules) in networks
SMART	Simple Modular Architecture Research Tool
SNP	Single Nucleotide Polymorphism
SP	Symbionts of plant roots and endophyte
SPFP	Symbiont of plant roots and endophyte with fungal pathogens
SwissProt	Manually annotated and reviewed subset of the UniProtKB database
TAPs	Transcription Associated Proteins
TrEMBL	Automatically annotated and not reviewed subset of UniProtKB database
UniProt	Universal Protein Resource
UniProtKB	Protein knowledgebase at UniProt
UniRef	UniProt Reference Clusters
WGS	Whole Genome Shotgun
WoLF PSORT	Program for protein subcellular localisation prediction
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 Overview

Numerous microbes retain the ability to invade a host and cause disease. In agriculture, disease outbreaks can have a significant impact on crop production and loss in harvest can lead to local or even global food and feed shortages.

The recent advances in the high-throughput technologies such as Next Generation Sequencing have led to the substantial increase of completely sequenced genomes for numerous plant pathogenic fungi (Aguilar-Pontes et al., 2014, Cantu et al., 2011, Cissé et al., 2013, DiGuistini et al., 2009, King et al., 2015). In order to assess, analyse, and discover vital information within these vast data sets, it is necessary either to develop new approaches or to use and modify existing approaches to be more efficient and thorough in their examination of the incoming data. This should permit scientists to reveal and predict the common and unique themes underpinning plant and animal pathogenesis, as well as to study the sequence variation responsible for differences in the biology of various strains of a single species.

Complex networks have become a valuable tool in the analysis, integration, and comparison of biological systems and their application in answering bioinformatics questions has increased considerably in the recent years (Altaf-Ul-Amin et al., 2014, Bennett et al., 2012, He et al., 2008, Liu et al., 2010, Lysenko et al., 2013, Zhao et al., 2009). Complex networks have been defined as the networks (graphs) with irregular structures that evolve dynamically in time (Boccaletti et al., 2006).

1.2 Research aims

The overall aim of this research is to explore, predict, and elucidate the unknown function of proteins within sequenced genomes / proteomes of economically important plant pathogenic fungi. The aim is further divided into four research goals:

- To identify plant pathogen-specific gene clusters and animal pathogen-specific gene clusters required for virulence, as well as those required by both pathogen types.
- To identify and predict the pathogenicity gene complement of two economically important plant pathogenic fungi, namely *Fusarium graminearum* and *Magnaporthe oryzae.*
- To investigate the role of Domains of Unknown Function (DUF) in the pathogenicity of the plant pathogenic fungus *Fusarium graminearum.*
- To perform a comparative network-based study between closely-related, as well as more distantly-related Ascomycetes, including both pathogenic and non-pathogenic fungi to reveal novel insights into pathogenicity.

1.3 Thesis outline

This thesis consists of eight chapters. Chapter 2 outlines fundamentals of protein structure and protein function prediction. It also provides an introduction to plant fungal pathogens biology and to the PHI-base (Urban et al., 2015a), as well as to network properties and their applications in biological systems such as protein-protein interactions. In particular, attention is given to protein-protein interaction network properties in order to find candidate pathogenicity genes in plant pathogenic fungi. Furthermore, previous work in the field is also described in Chapter 2.

Chapter 3 presents results from the general analysis of PHI-base version 3.2 content and reflects changes of the database content and size throughout the years of this study. In this chapter, clusters that contain genes associated with plants, animals, and both types of pathogens were identified. Identified genes associated with two plant pathogens, namely *Fusarium graminearum* and *Magnaporthe oryzae,* were implemented in Chapter 4 as 'seed' genes for prediction of candidate pathogenicity genes in these fungal species. In subsequent thesis chapters, different

versions of PHI-base were used to exploit the latest information from this growing resource over the lifetime of the thesis.

In Chapter 4, 'seed' genes information compiled in Chapter 3 is integrated using a protein-protein interaction network approach to predict candidate pathogenicity genes in two cereal-infecting fungal species, namely *Fusarium graminearum* and *Magnaporthe oryzae*. Moreover, characterisation of the *F. graminearum* genome based on functional categories of the genes within the *F. graminearum* genome is performed. Identification of the position of the predicted candidate pathogenic genes within the *F. graminearum* chromosome map is also presented. Finally, the work described and carried out in Chapter 4 initiated the co-authored publication (Lysenko et al., 2013).

In Chapter 5, the list of predicted candidate pathogenic genes taken from the Lysenko et al. (2013) study together with the information from PHI-base (versions 3.4 to 3.6) is used to explore the role of domains of unknown function (DUFs) in the disease-causing ability of the plant pathogenic fungus *F. graminearum*. Firstly, the pfam domain repertoire of the *F. graminearum* proteome is investigated with the main emphasis placed on the DUFs, their abundance within the proteome, and the location of the encoding genes within the four chromosomes. Furthermore, a taxonomic diversity evaluation of pfam domains and DUFs is presented for *F. graminearum.* Then, distinct domain-pair combinations (bigrams) are identified within the *F. graminearum* proteome. The bigrams are used further in the network analysis to examine the properties of DUFs and their possible impact on the pathogenic nature of *F. graminearum*.

Chapter 6 extends the methodologies developed in Chapter 5 to compare the pfam domains repertoire between plant pathogenic fungi and non-pathogenic fungi of *Fusarium* genus. The main emphasis is on the DUFs that were identified in Chapter 5 to be specific to fungal species.

In Chapter 7, the domain bigram approach, introduced in Chapter 5, is implemented to construct the Protein Bigrams Overlap Network (PBON) and Combined Protein Bigrams Overlap Network (CPBON). In this chapter a series of network analyses are attempted involving the following fungal species: *Fusarium graminearum*, *Fusarium culmorum*, *Fusarium venenatum* (non-pathogenic), *Magnaporthe oryzae*, and *Neurospora crassa* (non-pathogenic, model organism*).* Firstly, a PBON

is constructed for *Fusarium* species based on ordered-domains bigram-similarity. Then, PBON networks for three Ascomycete species: *F. graminearum*, *M. oryzae* and *N. crassa* are constructed based on ordered-domains bigram-similarity, whereas the connections between different networks are established based on orthologous proteins leading to CPBON creation. Furthermore, functional data obtained from the PHI-base and the extensive BROAD phenotyping platforms for *N. crassa* are applied to further explore the biology of the generated protein clusters.

Chapter 8 concludes the thesis and discusses the main findings and future directions of this research.

Chapter 2

Research background

2.1 Fundamentals of protein structure and protein function prediction

Proteins represent a large group of organic compounds consisting of amino acids linked by peptide bonds. Even though proteins are a relatively homogenous class of molecules, built of various combinations of 20 amino acids, they play a variety of functions in the organism. As structural proteins they provide the filamentous architecture of the cell. Proteins store and transport different sizes of particles ranging from molecules to electrons. Some control the passage of molecules across the membranes while others serve as enzymes to catalyse reactions within the living organism. As hormones, they pass the information between specific cells and organs in a complex organism. Proteins act as antibodies playing an important role in the immune system. Through the binding to specific sequences of nucleic acids, proteins control the gene expression, turning gene expression on and off. Proteins are also necessary for sight, hearing, touch, and other senses. The ability of proteins to perform such varied biological functions lies partly in the chemical diversity of amino acids and in the permutation of amino acids in the sequence, secondary modification such as glycosylation or phosphorylation, and the formation of three-dimensional structures that are stable in the normal protein environment.

2.1.1 Protein structure

There are four distinct levels of protein structure: primary, secondary, tertiary, and quaternary. The primary structure of a protein refers to the linear sequence of amino acids in the polypeptide chain connected by covalent peptide bonds. The ability to sequence amino acids was discovered by Frederick Sanger (Sanger, 1952). Each sequence of amino acids includes two ends: the amino terminus (N-terminal) and carboxyl terminus (C-terminal). The N-terminal end (NH2-group) indicates the start of the amino acid polypeptide, as NH2-group is the end where the amino group

is not involved in a peptide bond. The C-terminal end (a free carboxyl group (-COOH)) signifies the end of amino acid, where the -COOH-group does not participate in a peptide bond.

The secondary structure of a protein discloses the local three-dimensional conformation of the polypeptide chain imposed by the hydrogen bonds between backbone amino and carboxyl groups of the amino acids in the primary structure. The most common examples of the secondary structure are the alpha (α) helix, beta (β) sheets, loops, and turns. While secondary structures are local, the same protein molecule might comprise of many different secondary structures. Various combinations of connected secondary structure elements define the structural motif in the protein, or so-called 'super-secondary' structure of protein. Motifs are created by packing side chains from adjacent α helices and/ or β strands (sheets) close to each other. The tertiary structure of a protein is defined by the way the secondary structure elements of the protein are spatially positioned, relative to each other. The final three-dimensional tertiary structure of a protein is commonly described as a protein fold. Protein folding is a physical process by which a protein acquires its native functional shape or conformation. Several factors including, but not only limited to, non-specific hydrophobic interactions within the hydrophobic amino acids residues, specific tertiary interactions such as salt bridges, disulphide bonds, and packing the side chains are driving forces in the protein fold formation.

The tertiary structure of a protein illustrates the structural configuration of a single polypeptide chain. Despite this, fairly large numbers of proteins assembled from more than one polypeptide chain, known as protein subunits, associate into multiple-subunits representing the quaternary structure of a protein. The subunits of the quaternary structure can function either autonomously of each other or dependently so that the function of one subunit is driven by the function of other subunits.

2.1.2 Protein domains and superfamilies

Several motifs and secondary structure elements usually associate together forming a compact, self-stabilising, and semi-independent component called a **domain**, which often folds autonomously (Wetlaufer, 1973). Different but also overlapping definitions of the domain concept have been suggested since Wetlaufer's study in 1973. As a result of several independent

observations, domains are described as independent, self-folding, functional, and evolutional units of compact three-dimensional structures (Finn et al., 2014a, Murzin et al., 1995, Orengo et al., 1997, Riley and Labedan, 1997, Yeats and Orengo, 2001). Based on domains sequence and structure conservation the most commonly used frameworks to study protein evolution and the protein domains function are a superfamily and a family (Kotchoni et al., 2010). The superfamily concept was first introduced by Margaret Dayhoff's group in 1974 (Dayhoff, 1974, Dayhoff, 1976, Dayhoff et al., 1975) and resulted in Protein Information Resources (PIR) (Barker et al., 1993) which is still supported to the present day (Wu et al., 2003). However, Dayhoff's classification of protein sequences did not consider the presence of domains, was mainly based on a sequence similarity, and allowed a sequence to be assigned to a single superfamily.

In the mid-1990s, the domain concept became widely established with the emergence of the Structural Classification of Proteins (SCOP) (Murzin et al., 1995) and Class Architecture Topology Homologous superfamily (CATH) protein structure classification databases (Orengo et al., 1997). From that time superfamily resources such as SCOP, CATH, and PIR began to include the domain concept. Following on from the original protein superfamily definition, both SCOP and CATH categorised superfamilies based on similarities in a protein's sequence, structure and function. Related functions and similarities in a sequence or structure reflect their common evolutionary origin (Riley and Labedan, 1997). Thus, significant sequence similarity between protein sequences can be a powerful tool in identifying the members of protein families. Families tend to be clustered together into larger clades called superfamilies based on structural and mechanistic similarity, even though there is no detectable sequence similarity within the clustered sequences.

2.1.3 Proteins and protein domains classification resources

Numerous resources exist that classify proteins and protein domains to investigate the relationship between sequence, structure, and function. In brief, they can be divided into two groups, namely structure-based classification and sequence-based classification.

2.1.3.1 Structure-based classification

As previously mentioned, SCOP (Andreeva et al., 2008) and CATH (Sillitoe et al., 2015) are resources for structural classification of protein and protein domains. Both databases are organised on hierarchical levels.

The highest level in SCOP is represented by class which groups domains with similar secondary structures despite evolutionary origin or tertiary structure. Each class then consists of several folds, where the same fold indicates similarity in the tertiary structure but not necessarily evolutional relationship. Below the fold level, SCOP identifies a superfamily and a family layer. If domains share a distant common ancestor and perform a similar function, they are assigned to the same superfamily. However, sequence similarity between members of the same superfamily is very low. The domains within each superfamily are further assigned to a family if they share at least 30% of sequence identity or some sequence similarity if they perform the same function (Lo Conte et al., 2000).

The CATH database also defines a domain's hierarchical classification on four major levels: class, architecture, topology, and homologous superfamily; hence the name. At the class level, as per the SCOP database, domains are assigned based on their secondary structure. Then the architecture level distinguishes between different orientations of the secondary structures in three-dimensional space within the same class. However, it does not differentiate between different topologies (the connections between secondary structures). Structures which are grouped at the topology level have the same sequential connectivity of secondary elements but members of the same topology group might possess a diverse array of functions. Below the topology level, structures are grouped into homologous superfamilies. Each homologous superfamily comprises structurally and functionally similar protein domains. Domain sequences are assigned to the same superfamily if they share at least 35% sequence identity and have similar structure and function (Sillitoe et al., 2013, Sillitoe et al., 2015).

In view of the above, both SCOP and CATH use structure to define domain.

2.1.3.2 Sequence-based classification

The PFAM (the protein families database) is a sequence-based family resource that organises proteins and protein domain sequences into families. Each family is represented by a multiple sequence alignment and hidden Markov model (HMM) (Krogh et al., 1994). Members of the same family are expected to be functionally related sequences. However, as it was stressed in (Punta et al., 2012), functional annotation of the domain or protein cannot be based solely on the family membership. On the other hand, conservation of family signature residues or conservation of the common domain architecture might increase confidence in hypothetical function prediction (Punta et al., 2012).

PFAM entries are classified in one of four ways: family, domain, repeat, or motifs (Finn et al., 2014a). A domain in the PFAM is defined as an autonomous structural unit that can be found in multiple proteins. In contrast, a repeat is an unstable unit in isolation but forms a stable structure when present in multiple copies. Motifs describe a shorter sequence unit found on the outside globular domains (Bateman et al., 2002). Related PFAM entries are further catalogued together into clans. The relationship within proteins in a clan is defined by similarity of sequence, structure, or HMM profile. In contrast to SCOP and CATH, PFAM applies only sequence similarity to define a domain.

2.1.4 Protein function annotation

The concept of protein function can have many different features. Usually, protein function is determined by the molecular function of a sequence or a structure: catalytic activity of enzymes, transport and signalling activities of transmembrane proteins or scaffolding activity of structural proteins (Rentzsch and Orengo, 2009). However, protein function can also describe protein activity in the context of processes and pathways the protein takes part in, as well as the location where a certain molecular function takes place is supplementary to functional information of a protein.

Conventional schemes for the storage of information on molecular protein function include Enzyme Commission (EC), for enzyme protein, the Munich Information Center for Protein

Sequences (MIPS) Functional Catalogue (FunCat) (Ruepp et al., 2004), and the Kyoto Encyclopaedia of Genes and Genomes (KEGG).

The EC assigns a multiple digit E.C. number to a sequence encoding a specific enzymatic function. The E.C. number reflects the top-down hierarchy of enzyme function with the top representing the general enzyme classes. The Riley's scheme (Riley, 1993) and its successors the MIPS FunCat (Ruepp et al., 2004) and KEGG use a similar numbering system to EC to catalogue protein sequences into cellular processes or to specific conserved metabolic pathways respectively. The FunCat is a classification scheme that enables the description of proteins of prokaryotic and eukaryotic origin. It was developed initially for the *S. cerevisiae* genome project within MIPS. Since then the FunCat content has been significantly extended and it is used for a range of applications such as manual or automatic functional genome annotation, analysis of large-scale transcriptomic or proteomic data and has been extended to include additional organisms including plant pathogenic fungi, namely *F. graminearum* (see analysis performed in Chapter 4).

The most recent annotation system, Gene Ontology (GO) (Ashburner et al., 2000) assigns a function to gene and gene products across three categories: molecular function, biological process, and cellular compartment. Each category in GO is represented by a tree-like structure, the directed acyclic graph (DAG), in which GO terms are connected from the bottom to the top of the tree by child-parent relation. Each GO term can have multiple parent terms. In GO classification the most specific protein functions are represented by leaf nodes, whereas the root of the tree is assigned less specific molecular function, biological process, and cellular compartment.

2.1.4.1 *In silico* protein function annotation

Computational annotation of protein functions facilitates research into species which are less studied and therefore lack experimentally determined functional annotation. The simplest methods for inferring the molecular function focus on sequence similarity or homology search, the search for sequences encoding genes that share the common origin or ancestor (Koonin, 2005). Well-known examples of such an approach are Basic Local Alignment Tool (BLAST) and Position

Specific Iterative BLAST (PSI-BLAST) (for detection of remote homologues) (Altschul et al., 1997). Furthermore, the phylogenomic approach can expand accuracy of protein annotation further by differentiating orthologous and paralogous relatives.

Orthologs together with paralogs are two different types of homologous genes, the genes that share the common origin or ancestor. Orthologs are genes that evolved from the common ancestor genes by duplication in a speciation event leading to copies in different genomes. Paralogs are genes that evolved by duplication of the gene in the duplication event and can occur in both the same genome and different genomes. Usually, orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions[1] (Koonin, 2005, Walhout and Vidal, 2001). Furthermore, if interacting proteins X and Y in one organism have interacting ortholog proteins X' and Y' in another species, the pair of interactions X-Y and X'-Y' are defined as Interologs (Yu et al., 2004).

Both the sequence similarity and the phylogenomic approach rely to a large extent on the entire protein sequence similarity. Thus, following a sequence similarity check the protein function can be assigned to the complete gene's product (protein). However, a number of resources use protein domain or multi-domain combinations to predict the molecular function of the protein. Among these resources there are the SUPERFAMILY (Wilson et al., 2007a, Wilson et al., 2009), Gene3D (Lam et al., 2016), and PFAM (Finn et al., 2016) databases. All of them use HMMs to represent sequence families. SUPERFAMILY and Gene3D assign protein sequences to the structural domain families defined in the SCOP and CATH databases respectively, whereas PFAM assigns protein sequences and domains into families that share sequence similarities.

In contrast to the PFAM, SUPERFAMILY not only focuses primarily on SCOP superfamily level but also provides protein domain allocation at the family level (Gough, 2006, Oates et al., 2015). Consequently, PFAM, as a family-level database, might detect less distant relationships between

[1] http://homepage.usask.ca/~ctl271/857/def_homolog.shtml

domains and proteins but provides more specific information towards their function. Additionally, each domain in SUPERFAMILY, in contrast to PFAM, has a known 3D structure.

As a result of the above, we can observe a difference in family number between the SUPERFAMILY and PFAM resources, generally with a higher number of families in PFAM.

2.1.5 Domain combinations and versatility

The limited number of types of domains suggests that most domains are not exclusive to one protein but instead occur in a variety of proteins (Bork, 1991, Chothia, 1992). Moreover, the majority of unknown proteins can be mapped to known protein domains (Copley et al., 2002). As functional units of a protein, different domains are often associated with different functions. Most proteins, excluding a few disordered ones, consist of one or more domains that are clearly noticeable in a protein sequence, as well as in its three-dimensional structure (Apic et al., 2001a, Ekman et al., 2005, Murzin et al., 1995).

In the early 1970's, it became apparent that the same domain can reappear in various proteins with a different neighbour domain (Rossmann et al., 1974). Although most domains are found in the identical arrangement in proteins, there are highly versatile (or promiscuous) domains that form a variety of combinations with other domains. Furthermore, the distribution of the number of domain combinations for a given domain family, as well as the distribution of the number of different types of adjacent domains for each family follow a power law (Apic et al., 2001b, Wuchty, 2001). Thus, there is a small group of adjacent domains that dominate within each family and other types of domain combinations happen in the minority of proteins. Moreover, the majority of domain pairs appear only in one N- to C-terminal order or architecture, while only about 5% of domain pairs occur in both directions in Eukaryotes (Apic et al., 2001a). The above suggests that a new combination of domain pairs most likely results from a duplication of occurring combinations rather than recombination (Apic et al., 2003). Some domain pairs and domain triplets are conserved across domain architectures since they carry functions that can be adapted to a variety of domain contexts. These types of domains were labelled as supra-domains by Vogel (Vogel et al., 2004a, Vogel et al., 2004b).

As domain order in multi-domain proteins plays an important biological and evolutional role, the domain architecture in such proteins is often analysed in terms of adjacent domains pairs present in different proteins of a given genome. The adjacent pair of domains in the protein is often known as a 'domain combination' (Apic et al., 2001a, Vogel et al., 2004a) or 'bigram' (Basu et al., 2008, Seidl et al., 2011, Xie et al., 2011). The 'bigram' concept was originally derived from language modelling or speech recognition and refers to pairs of consecutive units such as letters, syllables, or words (Manning CaS, 1999).

Seidl et al. (2011) implemented the bigram concept into the study of the comparison of domain repertoire of 67 eukaryotic genomes including four oomycete and five fungal plant pathogens. The study confirmed an earlier finding by Basu et al. (2008) that the bigram numbers linearly depends on the number of domain types in the genome. In the Seidl et al. (2011) study, authors define each domain bigram as two successively emerged domains in a given protein. They also consider the order of domains with respect to N/C-terminus, so that bigram 'AB' is different from bigram 'BA'. Figure 2-1 summarises the approach used in the bigram identification and the domain abundance calculation implemented in the Seidl et al. (2011) study.

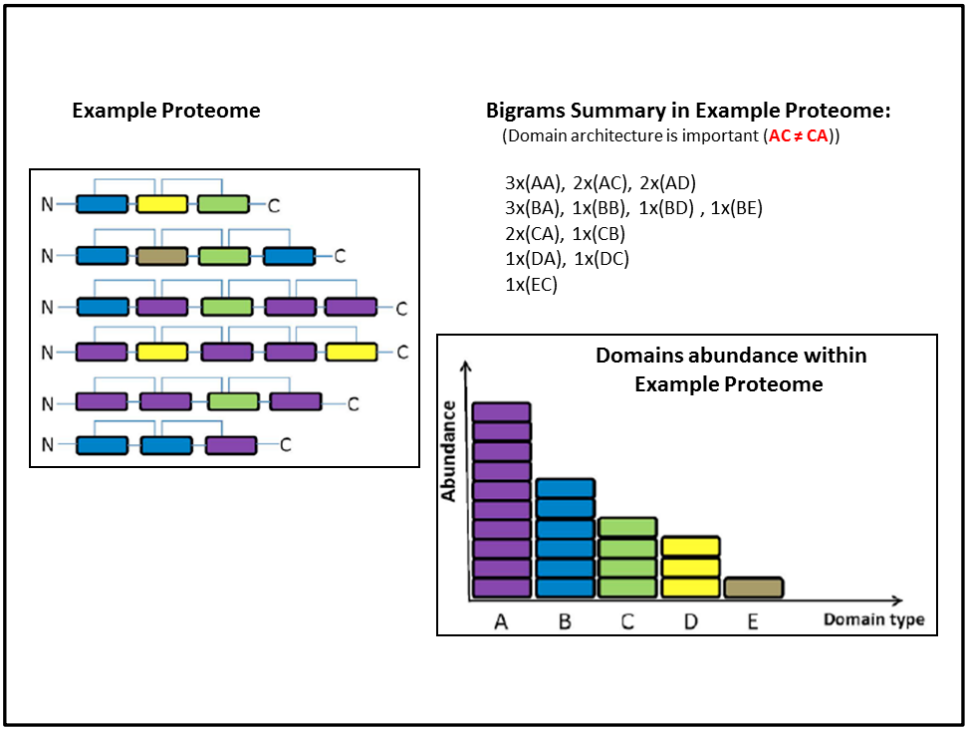


Figure 2-1 Different metrics used in Seidl et al. (2011) study.
Example proteome consists of five unique domains: A, B, C, D, and E.

In several studies the graph theory approach has been used to explore domain function and evolution across single or multiple organisms (Apic et al., 2001b, Apic et al., 2001a, Bork, 1991, Wuchty, 2001). Furthermore, Xie et al. (2011) employ a domain bigram network as a computational model for studying the cellular evolution across 77 eukaryotic genomes including fungal plant pathogen such as *F. graminearum*.

Commonly, a domain graph is represented as G (V_iV_j, E_{ij}), where nodes V_i and V_j indicate domains and an edge E_{ij} denotes the adjacency or co-occurrence of domains V_i and V_j within a given protein. Domain graphs are scale-free networks, indicating that the distribution of different types of adjacent domain numbers for each domain follows a power law (Wuchty, 2001) (please refer to section 2.3 of this chapter for a definition of power law (Equation 2-1) and scale-free network).

2.1.6 Domains of Unknown Function (DUFs)

Functional analysis of most newly sequenced genomes is carried out by transfer of annotation from similar sequences for which the function is known. However, there are still many proteins with no assigned function. Hypothetical proteins can be strict orphans, i.e. proteins that do not share homologues in genomes of other organisms and whose evolutionary origin is only poorly understood (Tautz and Domazet-Lošo, 2011). Alternatively, hypothetical proteins can belong to a family of related sequences of unknown function (Bateman et al., 2010). Thus, despite the intensive studies already carried out on sequenced genomes of both pathogenic and non-pathogenic species, there are still many regions within proteins, known as a Domain of Unknown Function (DUF), which need closer attention and further studies. The concept of DUF was firstly introduced by Chris Ponting, University of Oxford, through the addition of DUF1 and DUF2 to SMART (Simple Modular Architecture Research Tool) database (Schultz et al., 1998). Shortly after that, the proteins containing DUF1 and DUF2 were found to be involved in bacterial signalling pathways and in the processing of cyclic diguanylate. Consequently, their names were changed to GGDEF (PF00990, SM00267) and EAL (PF00563, SM00052) respectively. Both DUF1 and DUF2 were added to the PFAM database in 1997. Since then, new DUFs continued to be added to the PFAM and were no longer introduced to the SMART database (Bateman et al., 2010, Finn et al., 2014a, Punta et al., 2012).

DUFs are a large set of families within the PFAM. In the current database version 27.0 (at the time of writing the chapter), nearly 25% (3697) of domains are classified as DUFs (Finn et al., 2014a). If the function of at least one protein with a DUF has been experimentally resolved, the domain family is renamed and the given DUF ID is replaced by the family name. A study by Punta et al. (2012) demonstrates a substantial increase in the number of DUFs added to the PFAM over the last few years. Furthermore, the increase greatly overtakes the number of DUFs that have been functionally characterised and renamed. The authors of this study also highlight the fact that 23% of all DUFs within the PFAM (version 26.0) co-occurred with at least one other annotated domain and 76 of those DUFs were found in a single architecture in combination with at least one annotated domain. The term 'architecture' was defined by the authors as a combination of families occurring within the same protein sequence and in the study only architectures with at least five representative sequences were considered.

Assigning putative functions to DUFs will in turn help to determine a function for the protein or proteins containing each domain and thereby further our understanding of the proteome. There have been various attempts to catalogue these regions of protein sequences into families to enable understanding of the functional repertoire of unknown proteins. Jaroszewski et al. (2009) suggest that remote homology aided by the solved DUF structures (where available) could be used to assign likely function to unknown protein regions. Using this approach, DUF structure can be linked to a specific protein fold, as well as to a particular protein family with precise functional assignment. The study by Konc et al. (2013) developed an alternative method based on binding site comparison to propose the function of the unknown protein, and used this method to annotate DUF72.

In most cases, the function of each domain is the result of the synergistic relationship with other domains in the same protein and/or domains in interacting proteins (Vogel et al., 2004a). Seidl et al. (2011) predicted the domain repertoire of four oomycete plant pathogens and compared them with a wide variety of eukaryotes, including fungal plant pathogen such as *Fusarium graminearum* (FG) and various plant species. The number of unique pfam domains, as well as the number of different combinations of adjacent pfam domains, or bigrams, was calculated and used as a metric for the characterisation of the domain repertoire in several filamentous plant pathogens.

Additionally in the Seidl et al. (2011) study, differentially expressed genes during the infection were analysed which revealed a significant enrichment of genes with overrepresented domains in fungal and oomycete plant pathogens. Approximately 6% of the domains identified were found to be DUFs and were suggested to play an important role in the lifecycle of those pathogens.

2.2 Pathogens

Numerous pathogenic microorganisms have the ability to invade and cause diseases to various host organisms including human, animal and plant. Pathogens are described as infectious agents that are able to cause disease in its host in order to complete their life cycle (Shaner et al., 1992). Pathogenicity is the ability of a microorganism to produce an infectious disease in its host, whereas virulence (aggressiveness) indicates the relative degree of damage done by a pathogen (Baldwin et al., 2006). Plant pathogens have a significant impact on global agriculture and natural ecosystems because of the local, national, and international disease outbreaks which occur. Pathogens damage not only plants but also plant products on which humans are dependent for food (Fisher et al., 2012).

It is very important to understand the whole process that allows a pathogenic microbe to invade and infect their host. In general, the infection process consists of several distinct consecutive stages that lead to the development of a disease caused by pathogens. These primary stages are infective propagule arrival, penetration, establishment of infection, colonisation, growth and reproduction of the pathogen, dispersal of the pathogen, and survival of the pathogen without the host (Agrios, 2005) or within the host in the case of obligate biotroph pathogens (Knogge, 1996).

In the UK and Northern Europe, the vast majority of plant pathogens are represented by fungi and fungal-like organisms called oomycetes or protists. Both fungi and oomycetes have many similar features in growth, development, and plant-infection processes (Xu et al., 2006). Although most fungi and oomycetes are non-pathogenic, a few have a huge impact on agriculture and human welfare, by destroying or weakening important crops and some produce harmful mycotoxins (Marin et al., 2013). That is why these pathogenic species deserve focused attention from farmers, plant growers, plant advisors, and scientists alike.

The majority of phytopathogenic fungi belong to the Ascomycetes and the Basidiomycetes Phylum (Division). The destructive oomycetes belong to the Peronosporales, Pythiales, and Saprolegniales orders (Fawke et al., 2015).

Ascomycetes are commonly known as the Sac Fungi. This is because they form a microscopic sexual structure – 'ascus' in which ascospores form. Some species of Ascomycetes are asexual and do not reproduce via a sexual cycle nor form asci or ascospores (Agrios, 2005).

Basidiomycetes are filamentous fungi composed of hyphae (long branching filamentous cells). They reproduce in both asexual and sexual ways. During sexual reproduction, these fungi form specialised ‘club-shaped’ end-cells, called basidia, producing spores called basidiospores (Agrios, 2005).

Oomycetes are fungus-like eukaryotic microorganisms, also known as 'water mould'. Although they share some morphological characteristics with fungi, they exhibit several features which differentiate them from fungi. Oomycetes have a cell wall composed of cellulose not chitin and in the vegetative state they have diploid nuclei, whereas fungi are mostly haploid. They can reproduce in both asexual (zoospores) and sexual ways (oospores). Oomycetes comprise some of the harmful plant or animal pathogens (Fawke et al., 2015).

2.2.1.1 Classification of plant fungal pathogens by lifestyle

Most plant fungal pathogens can be classified as necrotrophs or biotrophs. Necrotrophs are the pathogens that kill their host cells in advance of the colonising hyphae and obtain the energy from the dead material (necrotrophy). Although these pathogens are characterised by wide host range, they exhibit weak to extreme virulence towards their host. The Ascomycete *Botrytis cinerea* is a typical plant pathogen representing the necrotrophs. This pathogen is able to infect at least 235 dicotyledonous plant species (Zhao et al., 2015).

On the contrary, biotrophs represent pathogens that are highly specialised towards one or a few particular hosts. The living host plant is absolutely necessary until the pathogen is ready to reproduce (Knogge, 1996). Thus, they obtain the necessary nutrients from the active metabolism of their host plant cells and they are completely dependent on the host organism as a source of

nutrients. Nearly all of the rust fungi (the cereal infecting pathogens *Puccinia spp*. and powdery mildew species, for example *Erysiphe graminis*) represent this type of pathogen. *Blumeria graminis* is a very good example of a biotroph. These species exist in two genetically different forms so-called ***formae specialis*** ('forms of the species'; f. sp.). One of the forms *Blumeria graminis* f. sp*. tritici* only infects wheat, whereas the other form *Blumeria graminis* f. sp*. hordei* only infects barley (Smith et al., 2010).

With regards to nutrition, some pathogens, such as the Ascomycete fungi from the genus *Colletotrichum* and Oomycetes from *Phytophthora* or *Peronospora* genus, employ first biotrophy, and subsequently, after the breakdown of host plant tissue, the hyphae switch to a necrotrophic growth phase. This kind of nutrition is called **hemibiotrophy** or **facultative biotrophy** (Koeck et al., 2011).

2.2.1.2 Plant immune system and fungal effector molecules

Most phytopathogenic fungi attack only a limited number of host species. However, some species, for example, *Botrytis cinerea* exhibit little specificity towards their host plants. *B. cinerea* can attack several plant species (Reis et al., 2005, Zhao et al., 2015) causing Grey-mould rot or Botrytis blight, and affects most fruits and vegetable crops, as well as a significant number of trees, shrubs, flowers, and weeds.

The different levels of specialisation of plant-pathogen interactions result from diversity in the host immune response and the pathogen's ability to evade or suppress host resistance (Figure 2-2). The primary plant immune response referred to as **PAMP-triggered immunity (PTI)**, where **PAMP** are **pathogen-associated molecular patterns**, is initiated when microorganism surface molecules such as cell-wall components like chitin or ergosterol are recognised by receptors present on plant cells (Chisholm et al., 2006, Postel and Kemmerling, 2009, Zipfel, 2009). Successful pathogens have evolved the ability to actively suppress the plant basal defence mechanism by secreting proteins or other compounds, known as **effector molecules,** into host cells. Such an occurrence is known as **effector-triggered susceptibility (ETS)** (Jones and Dangl, 2006). The secretion of effector molecules triggers a further plant defence mechanism towards the colonising pathogen resulting in **effector-triggered immunity (ETI)**. This takes place

in plants that have **resistance (*R*) genes** that encode receptors towards pathogen effector molecules. On the other hand, some pathogens are able to overcome ETI either by employing other effectors that suppress *R* gene-mediated defence in plants or by altering the recognised effector molecule (Bent and Mackey, 2007).

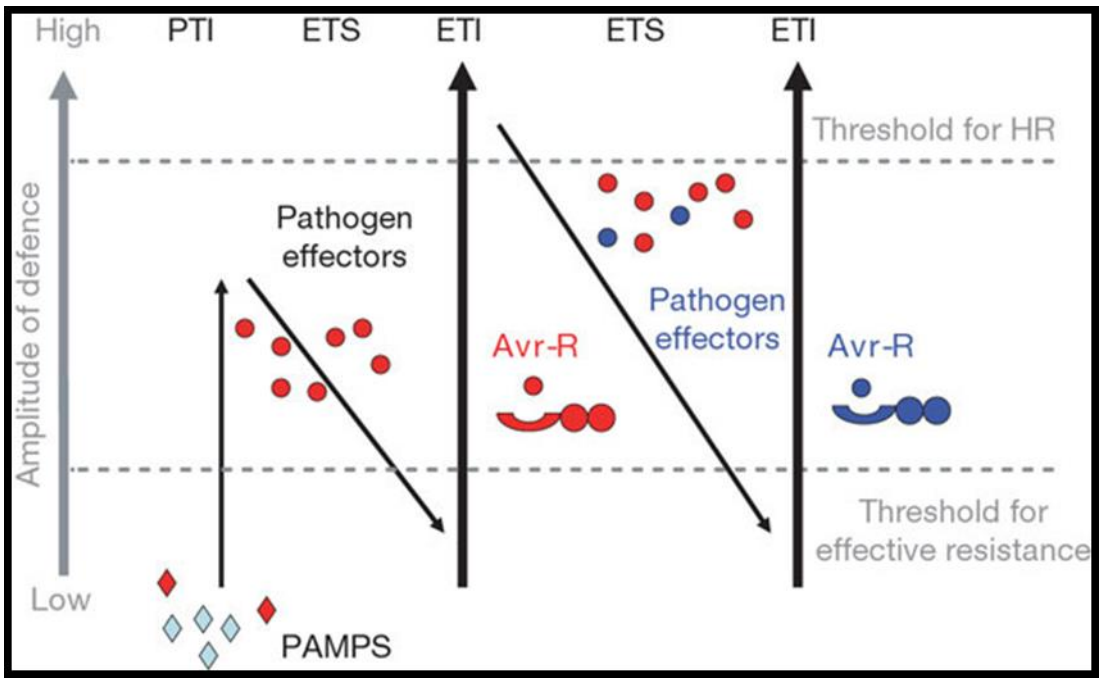


Figure 2-2 A 'zigzag' model of the quantitative output of the plant immune system.
Phase 1: plants detect pathogen-associated molecular patterns (PAMPs, red diamonds) via Pattern recognition receptors (PRRs) to trigger PAMP-triggered immunity (PTI). Phase 2: successful pathogens deliver effectors that interfere with PTI, or otherwise enable pathogen nutrition and dispersal, resulting in effector-triggered susceptibility. Phase 3: one effector (depicted in red) is recognised by R protein, activating effector-triggered immunity (ETI), an amplified version of PTI that often passes a threshold for induction of hypersensitive cell death. Phase 4: pathogen isolates are selected that have lost the red effector, and perhaps gained new effectors through horizontal gene flow (in blue)—these can help pathogens to suppress ETI. Selection favours new plant R proteins alleles that can recognise one of the newly acquired effectors, resulting again in ETI (Jones and Dangl, 2006).

2.2.1.3 Penetration into the plant tissue

To colonise a plant most phytopathogenic fungi have evolved various techniques for penetration into the host plant tissue. Fungi can attack different parts of the plant such as leaves, stems, or root. Regardless the part of the plant, the fungal pathogen can enter plants by three main routes: via natural openings such as stomata, penetration through intact surfaces, or entry through accidental wound sites. Some species can enter the plant via all routes, whereas the others are able to use only one route of entry.

Opportunistic pathogens penetrate the plant's surface through an accidental wound or attack plants, which are debilitated, stressed, or senescent (Brown et al., 2012b). Alternatively, some pathogens can enter the aerial parts of the plant, covered by a waxy cuticle, solely by producing hydrolytic enzymes, including cutinases, cellulases, pectinases, and proteases, causing digestion of the external barrier of the host plant.

Although *Botrytis cinerea* is thought to enter the host by producing extracellular enzymes (Doss, 1999, González et al., 2013, Zhao et al., 2015), other research has revealed the development of an appressorium-like structure as a way of penetration into the plant by this fungal species (Holz, 1995). However, the appressorium-like structure is not such a highly melanised **appressorium** as is formed by *Magnaporthe oryzae* (Howard and Valent, 1996, Jermy, 2012, Wang et al., 2005, Wilson and Talbot, 2009). The appressorium is firmly attached to the surface of the host. During appressorium development, huge turgor pressure is building up towards the host plant by the accumulation of the compatible solutes, especially glycerol, inside the structure. The porosity of the appressorium wall is significantly reduced by incorporation of melanin. The growing pressure is directed towards a small area at the base of the forming appressorium, which is free of wall material and melanin. Earlier studies showed that turgor pressure inside the appressorium acts as a driving force for penetration into the host epidermal cells (Howard et al., 1991). In some pathogens, the appressorium might have a penetration peg that pierces through the plant cell wall, whereas in other pathogens penetration involves mechanical perforation assisted by hydrolytic enzymes (Skamnioti and Gurr, 2007, Wang et al., 2005). Several studies demonstrated the importance of melanin in penetration via the appressorium (Chumley and Valent, 1990, Deising et al., 2000, Wolkow et al., 1983). Chumley and Valent (1990) proved that melanin deficient mutants in *Magnaporthe oryzae* were unable to infect the plants, whereas some of the mutants maintained the pathogenicity on leaves with the wounded epidermis. However, a recent study has demonstrated that in the anthracnose pathogen of corn, *Colletotrichum graminicola*, if turgor is generated whilst melanin biosynthesis is inhibited, the invasion of host leaves still occurs (Ludwig et al., 2013).

Many fungal species, not only pathogenic fungi, produce extracellular enzymes like pectinases, cellulases, cutinases, and proteases. Thus, those enzymes are unlikely to act individually as a

specially designed tool for pathogenicity, i.e. there is some redundancy in the system. However, it has been demonstrated that these enzymes, especially cutinases, play an important role during the active penetration of the pathogen throughout the plant wall during the infection process (Skamnioti and Gurr, 2008). Although the role of cutinases in fungal pathogens has been questioned (Stahl and Schäfer, 1992, Sweigard et al., 1992), there is strong evidence that disruption of cutinase gene CUT1 in *Nectria haematococca* (*Fusarium solani* f.sp. *pisi)* caused significant decrease in virulence of *N. haematococca* on pea plants (Rogers et al., 1994). Furthermore, a recent study of the role and the impact of cutinase on the pathogenicity of *Colletotrichum truncatum,* the chilli fruit pathogen, revealed the importance of cutinases in the pathogenicity process (Auyong et al., 2015).

The biotroph, the cereal powdery mildew fungus *Blumeria graminis,* penetrates the host with the aid of an appressorium and infection peg and then develops specialised feeding structures called **haustoria**. These are specialised branches of a fungal hypha formed inside a living cell of the host plant to obtain nutrients[2]. Although the haustoria penetrate the interior of the plant, both fungi and host plasma membranes stay intact. The necessary nutrients from the host travel via extracellular matrix which lies between both plasma membranes (Smith et al., 2010).

Thus, the penetration into plant tissue can range from entry through the abraded epidermis or a natural opening to various active penetrating methods through the outer surface of the plant.

2.2.2 Toxic secondary metabolites and mycotoxins

Following the penetration, to damage host-plant functions many non-biotrophic phytopathogenic fungi secrete toxic secondary metabolites. Some secondary metabolites are not only toxic to plants but also lead to huge economic losses in agriculture, because they also adversely affect animal and human health (Czembor et al., 2015, De Lucca, 2007, Marin et al., 2013). Mycotoxins that have a harmful effect on animal and human health are mainly located in post-harvest crops such as cereal grains. When food contaminated with mycotoxins is ingested by humans or animals, a wide variety of diseases, called mycoses, can occur with some leading even to death

[2] http://bugs.bio.usyd.edu.au/learning/resources/PlantPathology/glossary.html

or the formation of specific cancers. The other important route of exposure to mycotoxins is via the inhalation of spore-borne toxins during harvesting of grains (Bennett and Klich, 2003).

Mycotoxins are generally lipophilic and accumulate in the fat fraction of animal or plant (Gupta et al., 2011, Wild and Gong, 2010, Hussein and Brasel, 2001). However, some mycotoxins are water soluble and act upon specific host target proteins (Bennett and Klich, 2003).

Aflatoxins are a group of mycotoxins with the highest impact to human and animal health. Different species of the *Aspergillus* genus produce them. Although *Aspergillus* species are considered weak plant pathogens, species such as *Aspergillus flavus* (dominant on maize and cottonseed) and *Aspergillus parasiticus* (more common on peanuts) are among those producing very toxic, teratogenic, mutagenic and carcinogenic secondary metabolites called aflatoxins (De Lucca, 2007).There are four main aflatoxins produced by *Aspergillus* genus: B1, B2, G1, and G2. There is epidemiological evidence of an association between intake of aflatoxin B1 and liver cancer in humans (Hamed and Ali, 2013, Hifnawy et al., 2004, Kew, 2013, Wogan et al., 2004). There is also evidence that milk produced by cows that consume aflatoxin-contaminated feed contains an oxidative metabolic product of aflatoxin B1 – aflatoxin M1 (Nemati et al., 2010, Pathirana et al., 2010). Processing of infected grains leads to the release of airborne particles contaminated with aflatoxins. The exposure to those particles causes inflammation and irreversible pulmonary interstitial fibrosis in agricultural workers (Dvorackova and Pichova, 1986).

Fumonisins represent another important group of mycotoxins that play a significant role in agriculture, as well as in animal and human health. They are produced by *Fusarium* species, especially *Fusarium verticillioides, Fusarium oxysporum* and *Fusarium proliferatum* (De Lucca, 2007). Fumonisins are mainly found in maize and corn-based foods. Studies carried out by (Marasas et al., 2004, Merrill et al., 2001) suggest that these mycotoxins have significant disruptive effects on sphingolipid metabolism and can be potential risk factors for human neural tube development defects. There are two types of fumonisins: B_1 and B_2. Fumonisins B_1 is more toxic and is associated with the highest incidence of oesophageal cancer in humans (Bennett and Klich, 2003, Wild and Gong, 2010).

Trichothecenes are the other group of toxins produced by *Fusarium* species. There are type A and type B trichothecenes. Type A trichothecenes include T-2, HT-2, T-2 triol and T-2 tetraol. T-2 is the most important toxin of this group, since it inhibits eukaryotic protein synthesis, and it was found to be very toxic to leucocytes (Gutleb et al., 2002). Type B trichothecenes include deoxynivalenol (DON). The toxin is produced by *F. graminearum, F. culmorum* and *F. pseudograminearum* and is found in infected grains of corn, rice, oats, barley, and wheat. Humans are exposed to this toxin either directly by consuming contaminated plants such as grains, or indirectly via consumption of an animal-derived food like milk, eggs, liver or kidney (Sobrova et al., 2010). DON in human causes anorexia, nausea, vomiting, headache, abdominal pain, diarrhoea, chills, and convulsions (De Lucca, 2007).

Zearalenone (ZEA) is produced as a secondary metabolite by *Fusarium* species, mainly *Fusarium graminearum*, *Fusarium culmorum*, *Fusarium equiseti* and *Fusarium poae.* ZEA and its metabolites: α-zearalenol and β-zearalenol mimic estriadiol, interrupting physiological functions in humans or animals (Zielonka et al., 2015). The mycotoxin is often associated with reproductive disorders of farm animals and occasionally with hypoestrogenic disorders in humans (Hueza et al., 2014, Pistol et al., 2015, Zinedine et al., 2007).

2.2.3 The genomes of fungal species

The evolution in sequencing technologies has initiated a revolution in fungal and oomycete genomic analyses. Whole-genome shotgun (WGS) sequencing projects complementing Sanger sequencing have since been replaced with high-throughput Next Generation Sequencing (NGS) platforms such as Roche 454[3], the Illumina (Solexa)[4], the SOLiD sequencing system[5] and third-generation sequencing technologies such as Pacific Biosciences (PacBio)[6] system or the Ion Torrent Personal Genome Machine[7]. In addition, several assembly software algorithms have been developed for *de novo* assembly of NGS data (Miller et al., 2010). These include greedy

[3] http://allseq.com/knowledge-bank/sequencing-platforms/454-roche/
[4] http://www.illumina.com
[5] http://www.appliedbiosystems.com.
[6] https://www.pacb.com/
[7] https://www.thermofisher.com/order/catalog/product/4462921

algorithms such as: Short Sequence Assembly by progressive K-mer search and 3' read Extension (SSAKE) (Warren et al., 2007), Short read Assembler based on Robust Contig extension for Genome Sequencing (SHARCGS) (Dohm et al., 2007) and Verified Consensus Assembly by K-mer Extension (VCAKE) (Jeck et al., 2007), Overlap Layout Consensus Algorithms (OLC) including Celera Assembler with Best Overlap Graph (CABOG) (Denisov et al., 2008), as well as De Bruijn Graph-Based Algorithms: Velvet (Zerbino and Birney, 2008) or EULER-USR (Chaisson et al., 2009). This has led to a remarkable increase in the number of sequenced genomes available in the public domain and at a much lower cost.

Since the first fully sequenced eukaryotic genome became available for the fungus *Saccharomyces cerevisiae* genome (Goffeau et al., 1996), many members of the fungal genetics community have been actively working on providing additional fungal genomes via specific sequencing and genome-annotation projects. In 2000, a consortium of mycologists together with scientists from the Broad Institute (previously Whitehead Institute/MIT Center for Genome Research) launched the Fungal Genome Initiative (FGI)[8] to sequence genomes throughout the fungal kingdom. Since that time, several different fungal genomes have been released through the FGI and other centres including The Institute for Genomic Research (TIGR), Genome Sequencing Center at Washington University (WU-GSC), Sanger Institute and presently mainly by the U.S Department of Energy's Joint Genome Institute (JGI) and 1000 Fungal Genomes program[9] (Galagan et al., 2005, Grigoriev et al., 2011).

When starting this research in 2009, around 40 fully sequenced fungal genomes had been published in peer-reviewed articles, a similar number had been sequenced and were awaiting publication, and more sequencing projects were in progress. Figure 2-3 reproduced from the study by Aguilar-Pontes et al. (2014) illustrates the substantial increase in the number of sequenced fungal genomes in the last decade. Among the fully sequenced genomes available in the public domain, there are several genomes that represent important plant and / or human pathogenic fungi. Access to this genomic data is available by means of an ever-expanding number of online resources. These include the Comprehensive Phytopathogen Genomics

[8] https://www.broadinstitute.org/fungal-genome-initiative
[9] http://1000.fungalgenomes.org/home/

Resource (CPGR)[10], NCBI BioProject[11] 1000 Fungal Genomes program, and Genome OnLine Database (GOLD)[12] (Reddy et al., 2014).

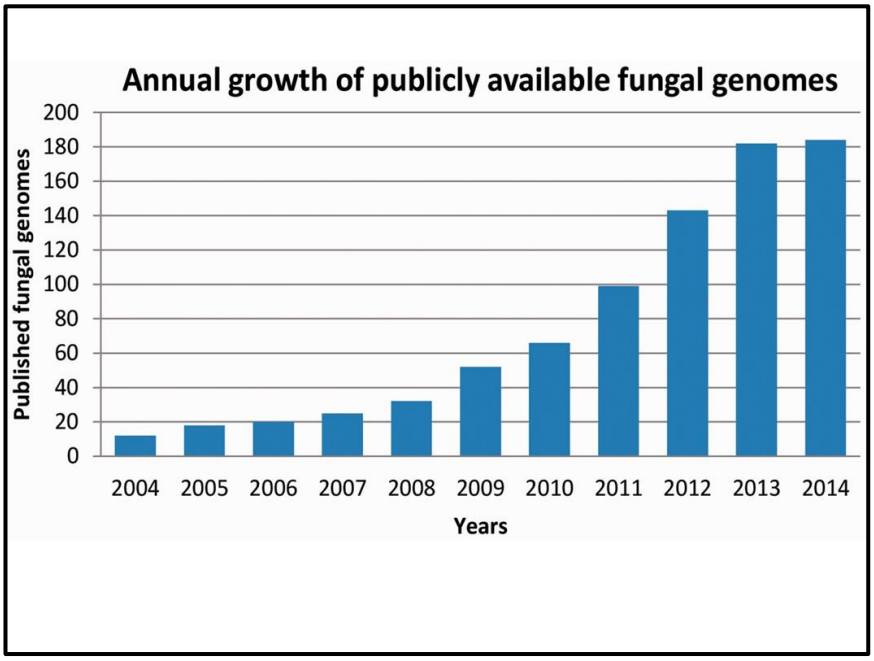


Figure 2-3 Annual growth of publicly available fungal genomes (Aguilar-Pontes et al., 2014).

2.2.4 Pathogenicity-related genes

The analysis of the growing numbers of complete pathogenic fungal genomes enables researchers to formulate hypotheses to predict the underlying biology of fungal infections and the pathogen-plant interactions, as well as the mechanisms by which fungi reproduce and stay in the environment. Furthermore, the growing numbers of fully sequenced fungal pathogen genomes and additional draft genomes of numerous isolates within a species may help researchers to discover a set of proteins that are unique to pathogenic fungi and/ or find the 'common core' within all pathogenic fungal species/isolates.

[10] http://cpgr.plantbiology.msu.edu/
[11] http://www.ncbi.nlm.nih.gov/bioproject/
[12] https://gold.jgi.doe.gov/

2.2.4.1 Cell wall and cutin degrading enzymes

Comparison of plant pathogenic fungi genomes (*Fusarium graminearum* and *Magnaporthe oryzae*) with non-pathogenic fungi genomes (*Neurospora crassa* and *Aspergillus nidulans*) revealed that there are certain gene families that are more highly represented in plant pathogenic fungi (Cuomo et al., 2007). In this early comparative study, genes coding for degradative enzymes were shown to be more abundant within pathogenic genomes. The genes encoding **cutinases,** cell-wall degrading enzymes important for penetration and colonisation of plant tissue, and **serine proteases** were reported in both plant pathogens (Auyong et al., 2015, Rogers et al., 1994, Skamnioti and Gurr, 2007).

The study by Cuomo et al. (2007) also showed that *F. graminearum* in contrast to other three fungi (*M. oryzae, N. crassa* and *A. nidulans)* has the highest content of **pectate lyases** (13) and pectin lyases (4), whereas *A. nidulans* has 8 of pectate lyases and 4 of pectin lyases as opposed to *M. oryzae* and *N. crassa* which has only 3 and 2 pectate-lyase genes respectively. Both saprophytic species lack pectin lyases. Formation of an appressorium (non-enzymatic penetration) by *M. oryzae* may explain the lower content of pectate lyase genes. It is worth emphasising that pectate lyases digest pectin - an essential component of plant cell walls. Later studies proved that genes encoding pectate lyase in *Colletotrichum gloeosporioides* (avocado fruits pathogen) (Miyara et al., 2008), *Colletotrichum coccodes* (the causal agent of black dot on potato and anthracnose on tomato) (Ben-Daniel et al., 2012), and *Colletotrichum lindemuthianum* (the causal agent of anthracnose in the common bean) (Cnossen-Fassoni et al., 2013) are required for virulence.

2.2.4.2 Mitogen-activated protein kinase (MAPK) genes

Genes encoding proteins from the highly-conserved MAPK pathways found in most eukaryotes have been shown to regulate various plant-infection processes, controlling fungal development, growth, and pathogenicity (Chen et al., 2014, Jenczmionka et al., 2003, Leng and Zhong, 2015, Urban et al., 2003). Two of three MAPK genes in *M. oryzae*: PMK1 and MPS1 regulate appressorium formation and penetration, respectively (Dean et al., 2005). A recent study (Penn

et al., 2015) revealed that chemical inhibition of PKC (kinase C) activities in *M. oryzae* led to a reduction in conidial germination, as well as appressorium development.

Similar to PMK1, the homologue in *F. graminearum* MAP1 (GPMK1) was found to be responsible for the signal-transduction process during the most important developmental stage in the life cycle of this pathogen. Disruption to this gene results in mutants that are sexually sterile and non-pathogenic due to a reduction in conidia production (Urban et al., 2003). Furthermore, two additional *F. graminearum* MAPKs, namely MGV1 and HOG1, were reported to play roles in female fertility and regulation of DON mycotoxin biosynthesis in *F. graminearum* (Jenczmionka et al., 2003, Zheng et al., 2012).

2.2.4.3 Genes coding for small effector proteins

Various non-enzymatic secreted proteins are likely to be crucial during fungal-plant interactions. Fungal plant pathogens secrete effector proteins to control, disable the host immune system and facilitate colonisation of fungi. Effectors are active outside the fungal cell and disturb the host defense process. They either activate or suppress plant-defense mechanisms. Most known fungal effectors are small and rich in cysteine residues proteins. *M. oryzae* appears to have twice the corresponding number of secreted proteins compared with *A. nidulans* or *N. crassa*. The novel variant cysteine pattern CX_7CCX_5C is widely expanded in *M. oryzae* (36 copies of 21 predicted proteins), whereas this pattern occurs only eight times in *A. nidulans* and four times in *N. crassa* (Dean et al., 2005). This represents a chitin-binding motif and may protect fungal cell walls from chitinases produced by the plant during its defense response to the attacking pathogen (Mentlak et al., 2012). However, a number of experimentally verified effectors have been reported and these are larger in size and not always rich in cysteines (Catanzariti et al., 2006, Djamei and Kahmann, 2012).

Several studies have incorporated specific characteristics into plant pathogenic fungal effectors (Rovenich et al., 2014, Mentlak et al., 2012, Marshall et al., 2011). It appears that there is a high level of diversity among fungal effectors, as well as low sequence similarity to other proteins, which indicates that the majority of effector proteins are species and / or clade-specific. This makes it difficult to anticipate their function in disease formation, and therefore necessitates a

thorough study of each type of putative effector proteins for a particular fungal pathogen. However, a recent study by Dong et al. (2014) revealed that a single amino acid polymorphism in the oomycete *Phytophthora* EPIC1 effector has been associated with pathogen ability to specialise to a new host.

2.2.4.4 Small molecule transporters

A comparative genomics study by Cuomo et al. (2007) also showed that the *F. graminearum* genome is rich in transporter genes for amino acids and sugar, as well as membrane-associated proteins that facilitate transport of several small molecules across the membranes. *F. graminearum* genome contains more predicted genes for transport facilitators as compared with *M. oryzae*, *A. nidulans* or *N. crassa*. Since specific transporters have been associated with toxins such as trichothecenes efflux (Gardiner et al., 2010), the higher number of major facilitator genes in *F. graminearum* indicates their potential importance either in the delivery of toxic metabolites into plants during the infection process or during the extensive saprophytic phase of colonisation on crop debris/soil surfaces.

2.2.4.5 'Common core genome' and mobile pathogenicity chromosomes

A study by Ma et al. (2010) on comparative genomics of phenotypically diverse *Fusarium* species, specifically *F. graminearum* (predominantly cereal host), *F. verticillioid*es (specific cereal host: maize), *F. oxysporum* f. sp. *lycopersici* (narrow species of non-cereal host: tomato), and *F. solani* (more diverged species to previous three), showed that each of the four *Fusarium* species carries a core genome with a high level of collinearity of chromosome segments (synteny) (Figure 2-4). However, *F. oxysporum* f. sp. *lycopersici* and *F. solani* each have lineage-specific (LS) chromosomes that are distinct with regards to repetitive sequences and genes related to host-pathogen interactions.

Additionally, the variation in LS chromosomes regions among *F. oxysporum* 5176 (Arabidopsis pathogen) and *F. oxysporum* f. sp. *vasinfectum* (cotton pathogen) strains revealed that there is a correlation between the different infection biology (host specificities) and pathogenicity-related genes on LS chromosomes.

A further experiment within the same study uncovered that, by mixing two strains on standard growth medium, transfer of two LS chromosomes by hyphal-tip fusion occurred between the *F. oxysporum* tomato-infecting strains, turning a nonpathogenic strain into a pathogenic one (Ma et al., 2010).

Advances in the analysis of entire genome for pathogenic fungi, determining the genes that are unique to pathogenic fungi, as well as genome-wide comparison will reveal further genes, metabolic pathways and chromosomal regions involved in pathogenicity. Comprehensive forward and reverse-genetics projects on gene disruption are currently underway for many economically important pathogenic fungal species. Collectively, these new data sets will enable a systems approach to understand the biology underlying the widespread agricultural and medical impact of pathogenic fungi.

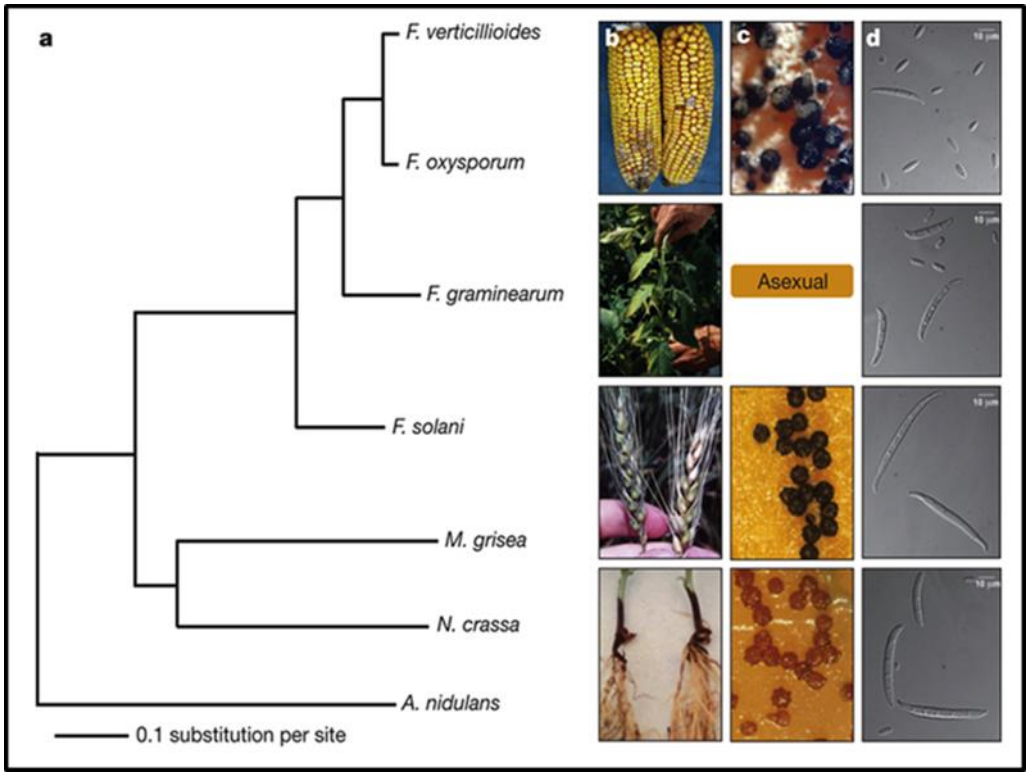


Figure 2-4 Phylogenetic relationship of four *Fusarium* species in relation to other ascomycete fungi and phenotypic variation among the four *Fusarium* species.
a: Maximum-likelihood tree using concatenated protein sequences of 100 genes randomly selected from 4,694 *Fusarium* orthologous genes that have clear 1:1:1:1 correlation among the *Fusarium* genomes and have unique matches in *Magnaporthe oryzae*, *Neurospora crassa* and *Aspergillus nidulans*. b: Disease symptoms of (top to bottom) kernel rot of maize (*F. verticillioides*), wilt of tomato *(F. oxysporum* f. sp. *Lycopersici*) head blight of wheat (*F. graminearum*) and root rot of pea (*F. solani*); c: The perithecial states of *F. verticillioides*, *F. oxysporum* f. sp. *Lycopersici* (no sexual state), *F. graminearum* and *F. solani* and d: Micro- and macroconidia of *Fusarium* species respectively. Scale bars: 10 µm (Ma et al., 2010).

2.3 The Pathogen-Host Interactions database - PHI-base

With the advent of new sequencing technologies, completely sequenced genomes are becoming available for hundreds of pathogenic species (see section 2.2.3). Over the years, the number of publications on gene sequence and functional information on pathogenicity, virulence, and effectors genes for many species, including bacteria, viruses, oomycete and fungi, have grown considerably (see Table 2-1). For many scientists, it is often difficult to access full data sets and evaluate the results because such information is generally accessible either via searching the primary literature or in the laboratories of individual scientists.

The demand for web-accessible resources that collect, cross-link, and catalogue genotypic and phenotypic information on particular pathogens and gene deletion mutants was fulfilled by the official launch of PHI-base[13] in 2005 (Winnenburg et al., 2006). Although other databases such as PathoPlant[14] (Bolivar et al., 2014) are available, PHI-base is a unique database in that it focuses on genes involved in pathogen-host interactions, and gene functions that are experimentally verified. This makes PHI-base a valuable resource in the discovery of genes that play an important role in human and animal health and in agriculture.

Each entry in PHI-base is manually curated by a number of experts and supported by strong experimental evidence (gene disruption, gene silencing or other alteration experiments), as well as literature references in which each experiment is described. Thus, each entry of PHI-base is an experimentally verified pathogenicity, virulence or effector gene from fungal, bacteria, oomycete or other protist pathogens, which infect human, plant, animal, fish, fungal and insect hosts. The genes are labelled as pathogenicity genes if the effect on the phenotype has a qualitative outcome (disease or no disease). Analogically, the genes are labelled as virulence genes if the effect on the phenotype is quantitative. Effector genes represent another group catalogued genes in PHI-base.

[13] http://www.phi-base.org.
[14] http://www.pathoplant.de.

Each gene in the database is presented with its nucleotide and corresponding amino acid sequence, as well as a detailed description of the predicted protein function during the host-infection process. For each gene, one or several interactions with a host is assigned. In other words, an interaction links a gene with one host and one tissue type (Urban et al., 2015b).

In the PHI-base version 3.8, nine phenotyping terms are used to describe the phenotype outcome for one interaction: loss of pathogenicity, reduced virulence, unaffected pathogenicity, increased virulence, effector gene (plant avirulence determinant), lethal, enhanced antagonism, resistant to chemical, sensitive to chemical (Urban et al., 2015b). Of interest here, there are five phenotyping terms, namely 'loss of pathogenicity', 'reduced virulence', 'unaffected pathogenicity', 'increased virulence', and 'lethal'. 'Loss of pathogenicity' describes a gene mutant that fails to cause disease in contrast to the gene in the wide-type strain. Similarly, the 'reduced virulence' outcome suggests that the gene mutant still leads to disease but with fewer symptoms than the wild-type strain.

The 'unaffected pathogenicity' term indicates that the gene alone does not play the role in a pathogenic process under the investigation and /or that genetic redundancy could be operating. However, the inclusion of double-gene deletion for example might have an opposite or different phenotypic outcome compared to the single gene studies (Wilson et al., 2007b, Zhang et al., 2011). The 'increased virulence' phenotype outcome indicates that the mutant of the gene causes greater occurrence and severity of disease than in the wide-type strain (Gardiner et al., 2009). Thus, increase virulence outcome is an example of negative regulation of key pathogenicity processes taking places during the infection of either plants or animal hosts (Brown et al., 2015). Finally, the term 'lethal' indicates that the gene is essential for the life of the tested organism (Wang et al., 2011a).

PHI-base can be used not only to find the original publication, or to undertake a keyword search, but also enables direct access to other relevant literature, information on the studied gene,

protein, GO terms or interaction via a direct link to UniProtKB (Protein Knowledgebase)[15] (Consortium, 2015) and PubMed[16].

Additionally, PHI-base phenotypes are mapped to Ensembl Genomes[17] (Kersey et al., 2014) sites for fungi, protist (including oomycetes), and bacteria via their genome accession (Urban et al., 2015b). Since 2011, phenotype information from PHI-base is also accessible via plant pathogen genome browsers at PhytoPath database[18] (Pedro et al., 2016).

Table 2-1 Changes in the PHI-base content within the six years of study.

Release date	Dec 2009 - Jan 2012	Feb-13	Nov-13	May-14	May-15	Jul-15	Sep-15
Version	3.1 to 3.3	3.4	3.5	3.6	3.7	3.8	4.0
Genes	1065	2433	2434	2875	3369	3562	3944
Interactions	1335	3151	3152	4102	4792	5017	6473
Pathogens	97	109	109	166	225	231	262
Hosts	76	92	92	110	132	143	165
Diseases	64	107	108	181	261	283	374
References	720	1049	1049	1243	1693	1812	1881

2.3.1 Application of PHI-base

PHI-base, a freely available resource to both academic and non-academic organisations, is directed towards a wide group of users. This includes scientists from medical and agricultural disciplines, and it is also a valuable resource for bioinformaticians and evolutionary biologists, providing an easy access to peer-reviewed data on a wide range of genes from pathogen species that invade plant, animal, fungal and insect hosts. The possibility of downloading all nucleotide or protein sequences present in the database in FASTA format allows users to identify known pathogenicity genes in newly sequenced genomes, and also perform comparative analyses, genomic or transcriptomic studies, proteome annotation and other predictive bioinformatics analyses (Agrawal et al., 2015, Liu et al., 2010, Lysenko et al., 2013, Schleker et al., 2012, Sperschneider et al., 2013, Thakur et al., 2013, Vargas et al., 2012, Zhang et al., 2014).

[15] http://www.uniprot.org
[16] http://www.ncbi.nlm.nih.gov/pubmed/
[17] http://ensemblgenomes.org/
[18] http://www.phytopathdb.org/

In addition, the whole content of the database can be downloaded in Extensible Markup Language (XML). This makes it possible to use and integrate PHI-base in other external applications such as the ONDEX system (Köhler et al., 2006).

Furthermore, the current version of PHI-base includes genes that are verified targets of known bioactive compounds which kill pathogens or arrest their growth. Comparisons of sequence similarity make it possible to infer, across species, which bioactive compounds can also be used as anti-infectives against new pathogens. To support this type of comparative analysis fungicides and target genes catalogued by Fungicide Resistance Action Committee (FRAC)[19] have been included in PHI-base. Due to requests from the global community, when a gene sequence has been found not to be required for pathogenicity or virulence this negative information is also added into PHI-base.

Finally, the growing number of publications citing PHI-base and reaching 122 scientific references in August 2015 (please see the up-to-date number of publication under “About” section of PHI-base website) indicates the popularity of PHI-base (Urban et al., 2015a).

2.4 Complex networks

The world surrounding us is full of networks. Cellular phones, the electric power grids, the world-wide internet network, and highway and railway systems are important elements of our everyday life. Even we as individuals are part of a social relationships network.

One of the most broadly studied class of networks in the literature is a biological networks class (Bennett et al., 2015, Durmuş et al., 2015, Liu et al., 2010, Stuart et al., 2003, Wang et al., 2011b, Wu et al., 2006, Xia et al., 2010, Zhao et al., 2009). The class comprises of multifunctional networks representing biological processes in the single or coherent context (de Silva and Stumpf, 2005). Within this class we can distinguish transcriptional (Sorrells and Johnson, 2015), metabolic

[19] http://www.frac.info/.

(Ma and Zeng, 2003, Radrich et al., 2010) and protein-protein interaction networks (Lysenko et al., 2013, Zhao et al., 2009). These networks are characterised by considerable topological features where patterns of adjacency between vertices in the network are neither regular nor random (Albert and Barabási, 2002, Boccaletti et al., 2006).

In mathematical language, a network is a graph consisting of vertices (nodes), which signify the elements or objects of the network, joined together by links (edges), which show relationship between the elements (Boccaletti et al., 2006, Buchanan M. et al., 2010, Cohen R., 2010). Throughout this work, the terms vertices and nodes, as well the terms links and edges, are used mutually. In general, we can distinguish directed and undirected graphs. In **directed graphs** (partially depicted in Figure 2-5), the edges are considered as ordered pairs of nodes and each edge has a direction pointing from the first to the second node in the pair. Analogically, in the **undirected graphs** (partially depicted in Figure 2-5) the edges are not ordered pairs of nodes. Figure 2-5 was created based on the information in (Boccaletti et al., 2006, Buchanan et al., 2010, Cohen, 2010). For example, link (1,2) presented in Figure 2-5 indicates that node 1 is either a source or a target, as well as that node 2 is either a source or a target in the pair of nodes 1 and 2. On the other hand, link (2,4), depicted in that figure, indicates that node 2 is a source and node 4 is a target in the pair of nodes 2 and 4.

Gene regulatory network is an example where biological systems (a group of structures that work together to perform a specific task) can be efficiently modelled by implementation of a directed graph. In such a network, nodes represent genes and directed edges denote direction from transcription factor (the product of the first gene in the pair) to the second gene in the pair which it regulates. A protein–protein interaction (PPI) network, however, is an example of undirected network application in biological systems. Nodes in such a network represent proteins, and edges are physical mutual interactions between these proteins.

If there is a path in the undirected graph connecting the nodes through a finite number of links, we can say that those nodes belong to the one **connected component** of the graph. The number of nodes determines the size of the connected component. The larger the connected component in the network, the more efficient is the spread of the information within the network (Boccaletti et al., 2006).

In order to understand the properties of the network, it is necessary to define metrics to characterise its topology. The main measurements are node degree, degree distribution, clustering coefficient (transitivity), shortest distance, diameter, betweenness, and closeness (Boccaletti et al., 2006).

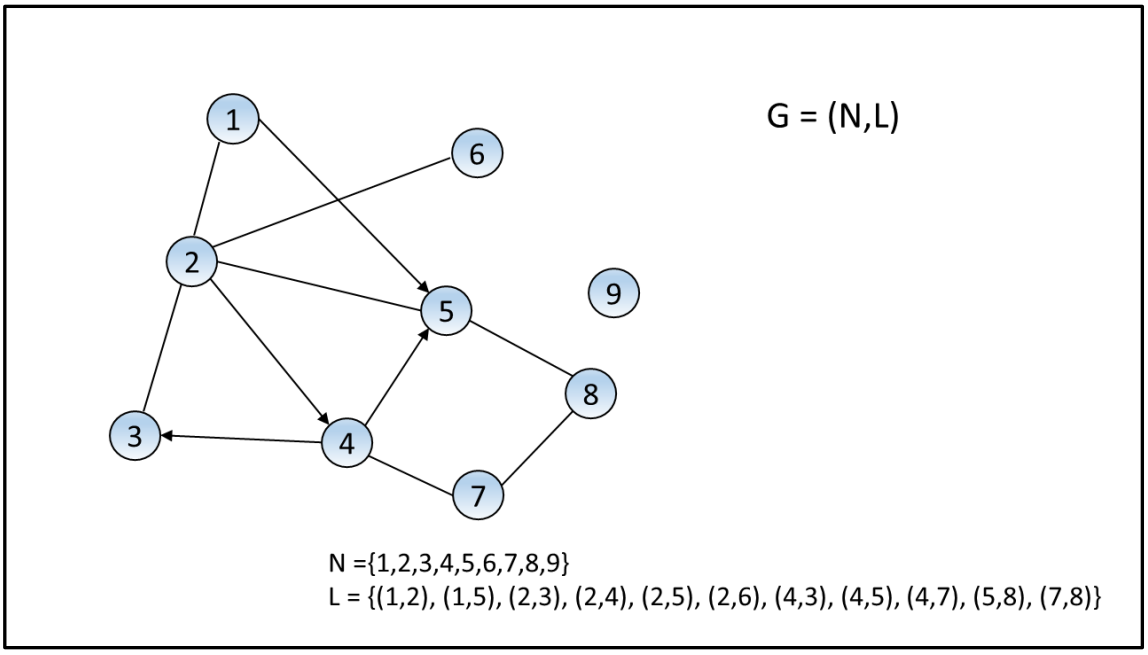


Figure 2-5 Partially directed graph (network).
G is a partially directed graph (network) where N is a set of nodes in the graph and L is a set of links in the graph. Labelled nodes and links are respectively represented by circles and straight lines in undirected relation (undirected graph) or arrows in directed relation (directed graph).

Node degree, *k,* in the network is the number of edges connected to the node, whereas **degree distribution, *P(k)*** represents the fraction of nodes in the network with degree k. In other words, it is a probability that randomly selected node in a given network has a degree *k.* The degree distribution is also the simplest measure of a network, and it is commonly the first step in describing the network (Boccaletti et al., 2006). Complex networks such as biological systems like PPI networks are characterised by degree distribution that follows a power law (Equation 2-1) (Barabasi and Albert, 1999, Cohen R., 2010).

Equation 2-1

$$P(k) \sim k^{-\gamma}$$

Consequently, these networks are known as **scale-free networks** with a small number of highly connected vertices (the so-called 'hubs' nodes with high vertex degree) and most vertices with comparatively few links (Barabasi and Albert, 1999). Thanks to their scale-free distribution, complex networks exhibit a remarkable degree of robustness to perturbations in communication between their nodes. The robustness is important in biological systems and makes them more resilient to random mutations. On the other hand, networks are vulnerable to targeted attacks on hubs, leading to drastic changes to the network topology (Albert et al., 2000, Jeong et al., 2001).

Another important property of a node in the network is the node **clustering coefficient** (Figure 2-6) and its global measure **average clustering coefficient** of the network, known as network transitivity (Boccaletti et al., 2006). Figure 2-6 was made based on the image in (Buchanan et al., 2010). If the link marked with the dashed line exists, the clustering coefficient of node *i* would be equal to 1. Otherwise, the clustering coefficient of node *i* is equal to $^{2}/_{3}$.

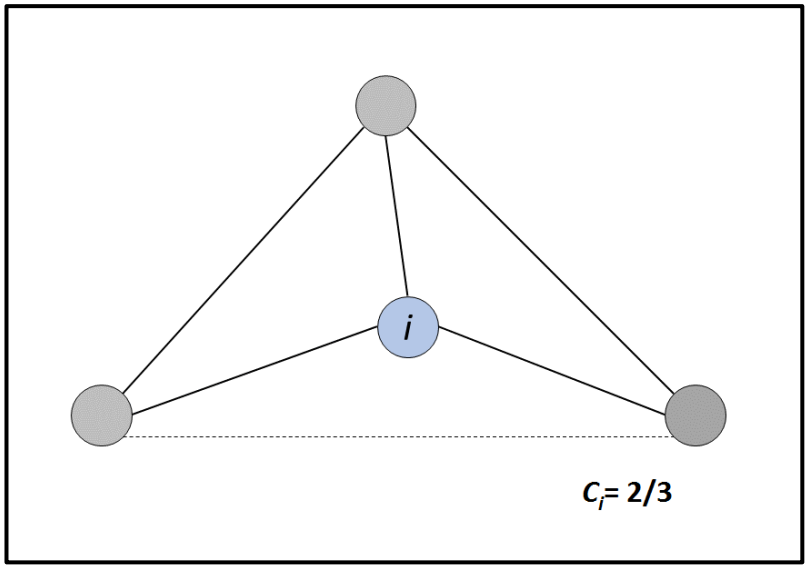


Figure 2-6 Clustering coefficient (C_i) of node i.
Nodes and links are respectively depicted by circles and straight lines. The dashed line indicates an absence of the link between given nodes.

The clustering coefficient in an undirected graph represents the number of links joining the neighbours of node *i* to each other, divided by the possible number of neighbours pairs of node *i.* In other words, clustering coefficient determines the degree to which the nodes in the graph tend to cluster together (Cohen, 2010, Girvan and Newman, 2002). The clustering coefficient C_i for a given node *i* in an undirected graph can be computed with the aid of Equation 2-2 (Boccaletti et

al., 2006), where k_i is the number of neighbours of node *i* and e_i is the number of connected pairs between all neighbours of node *i*. Computing the average clustering coefficient, C, over all nodes in the network reveals the probability of finding the link between two randomly chosen neighbours of a given node in the network.

Equation 2-2

$$C_i = \frac{2e_i}{(k_i\,(k_i - 1))}$$

Another important property which plays a significant role in communication and transport within the network is the **shortest path** (shortest distance) between nodes (Figure 2-7). The figure was redrawn from Buchanan et al. (2010). For each pair of nodes *i* and *j* in undirected graphs, the shortest path length d_{ij} is the minimum number of links that must be crossed to go from node *i* to node *j* (Boccaletti et al., 2006). Thus, the shortest path length highlighted in red in the Figure 2-7 is equal to 3.

The largest value of the shortest path between nodes in the whole network is defined as the **diameter** of the network. It is also possible to calculate the **average distance** over all pairs of nodes. However, such distance should be calculated for the nodes within the connected component to avoid infinitive value of the average distance. Both diameter and average path length are important in characterising the **small-world property** of the network. We deal with the small-world property of the network when the average path length and diameter are very small despite of the large number of nodes in the network (Buchanan et al., 2010, Telesford et al., 2011).

Another measurement of node importance in the graph is its **betweenness centrality** (Figure 2-8). The betweenness of node *k* in a network is equal to the number of shortest paths from all vertices to all others vertices that pass through it (Buchanan et al., 2010). Thus, the betweenness centrality of node *k*, depicted in Figure 2-8, is computed considering all shortest paths between nodes *i* and *j* that cross through node *k*. For nodes *i* and *j* in the figure there are two possible shortest paths and each of them contributes to the weight of ½. Thus, betweenness centrality of

node *k* equals to ½, as only one of two shortest paths between nodes *i* and *j* crosses through node *k*.

The centrality of node *k* increases with the growing number of the shortest paths crossed through node *k*. The betweenness centrality of node *k* can be calculated from Equation 2-3 (Buchanan et al., 2010).

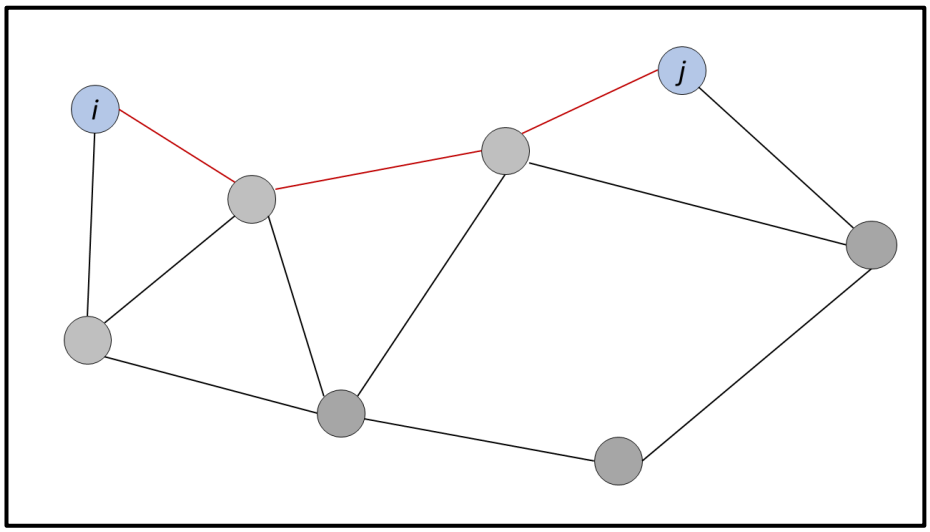


Figure 2-7 The shortest path length between two nodes in the network.
Nodes and links are respectively depicted by circles and straight lines. The red links indicate the shortest path length between nodes *i* and *j*.

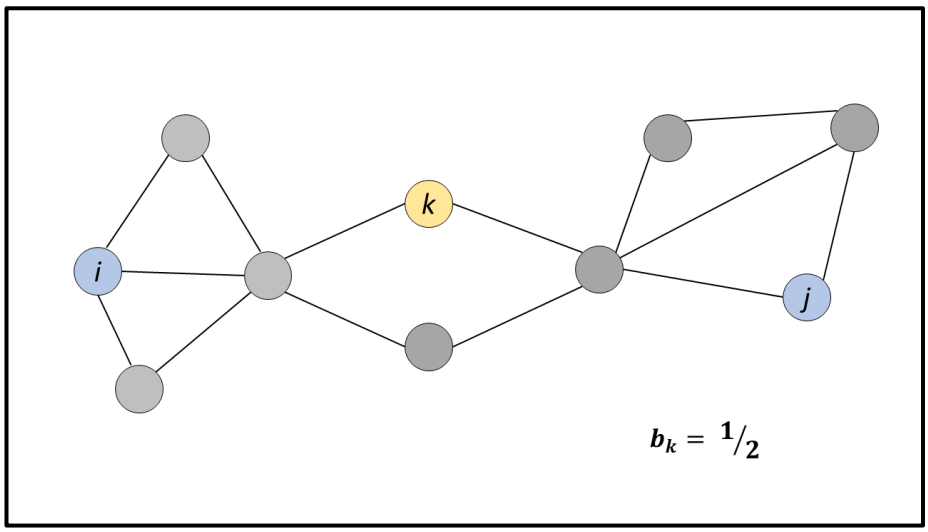


Figure 2-8 Betweenness centrality of node k.
Nodes and links are respectively depicted by circles and straight lines. Here, the betweenness centrality of node k is computed by considering all shortest paths, crossing through node k, between nodes *i* and *j*.

Equation 2-3

$$b_k = \sum_{i \neq j \neq k} \frac{N_{ij}\,(k)}{N_{ij}}$$

Where the sum runs from 1 to *N* (total numbers of nodes) and node *i* is different from node *j* and both are different from node *k*. The value of N_{ij} indicates the total number of shortest paths between *i* and *j* nodes, whereas $N_{ij}\,(k)$ is the number of shortest paths that cross through node *k* (Buchanan et al., 2010).

A further measure of a node centrality in the graph is closeness centrality. **Closeness centrality** is defined as the inverse of the sum of all shortest paths to other nodes in the network or the smallest number of ties to go through to reach all other nodes individually (Opsahl et al., 2010).

As indicated earlier, complex networks have a scale-free distribution of nodes. In other words, in the real network we have a topology where nodes with low degree coexist with nodes with large degree. This also applies to the edge distributions in the real networks where the density of edges within particular groups of nodes is higher than the average edge density in the whole network. Such groups of nodes with a high density of edges within them are defined as **community structures** (also known as modules or clusters) (Fortunato, 2010, Fortunato and Castellano, 2012), see Figure 2-9 reproduced from Fortunato and Castellano (2012). Each community consists of nodes that share similar properties or play a similar function in the graph. Thus, in protein-protein interaction networks proteins that are within the same community are likely to share the same specific role within the cell (Fortunato, 2010).

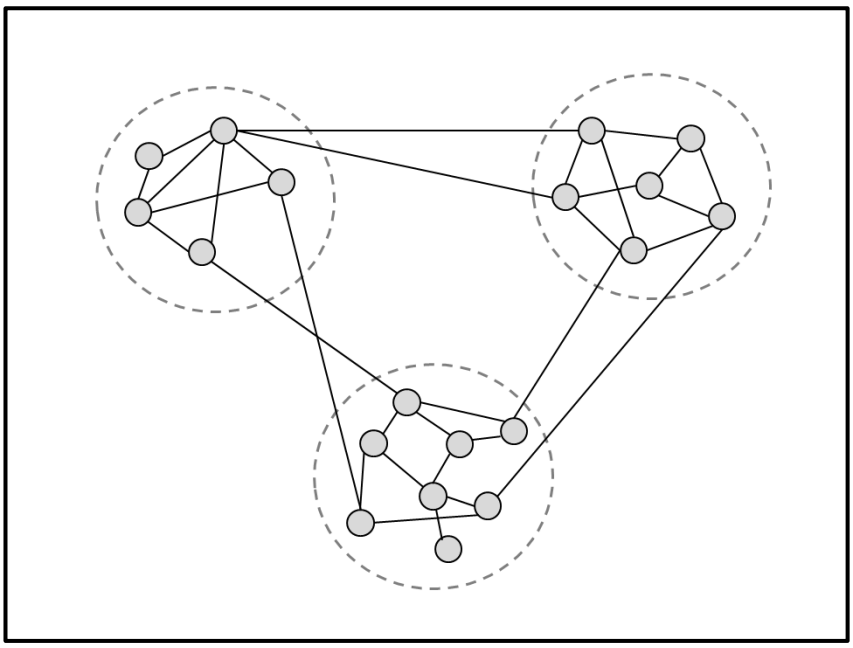


Figure 2-9 A simple graph with three communities.
Nodes and links are respectively depicted by circles and straight lines. Each dashed circle represents community structure (module).

2.5 Application of networks in plant-fungal interactions

Biological networks, such as protein-protein interaction (PPI), metabolic, or transcriptional networks, have become one of the most well-studied classes of networks in the literature (Bennett et al., 2015, Durmuş et al., 2015, Liu et al., 2010, Lysenko et al., 2013, Ma and Zeng, 2003, Radrich et al., 2010, Sorrells and Johnson, 2015, Stuart et al., 2003, Wu et al., 2006, Xia et al., 2010). This section focuses on the PPI network and its application in the prediction of candidate pathogenicity genes in plant pathogenic fungi.

Pathogenicity is a complex process that is known to involve many molecular regulations and interactions (previously described in sections 2.2-2.3). This is because the majority of proteins carry out their function by interacting together in a complex manner as a part of a network.

Recently large-scale data relating to protein-protein interactions has become available for several organisms (Orchard et al., 2014). For those organisms where no experimental data are available,

it is sometimes possible to predict candidate genes that take part in the protein-protein interactions in other organisms based on orthology relationships (see section 2.1.4.1).

A PPI network is a graph where the nodes represent proteins and the edges denote interactions. Thus, having a PPI network for plant pathogens offers the possibility of exploring the network neighbourhood of known pathogenicity-associated proteins and making predictions.

Currently, there are limited experimental data available on globally surveyed PPI in plant pathogenic fungi. Also, there are few predicted interactomes (protein-protein interactions) for plant pathogenic fungi. For the purpose of this work, the interactomes for two economically important plant pathogenic fungi, namely *Fusarium graminearum* (Zhao et al., 2009) and *Magnaporthe oryzae* (He et al., 2008) are of the main interest here.

2.5.1 General characteristic of data available for plant pathogenic fungi

Fusarium graminearum (teleomorph *Gibberella zeae*) is a plant pathogen which causes Fusarium Ear Blight (FEB), also known as Fusarium Head Blight, or Fusarium Scab – a devastating disease of wheat, barley, oats, rye, and maize crops globally. Additionally, mycotoxins such as DON produced by this pathogenic fungus during infection make the harvest grains unsafe for consumption (Goswami and Kistler, 2004).

2.5.1.1 *Fusarium graminearum* genome

Sequence data

With the advent of Sanger sequencing technology, the genome sequence for the PH-1 strain of *Fusarium graminearum* was first available in 2003 and then annotated at the Broad Institute and the global *Fusarium* / fungal community (Cuomo et al., 2007). Subsequently, the second gene call for this destructive pathogen became shortly available afterwards as a result of the FGDB (*Fusarium graminearum* Genome Database) project within MIPS (Munich Information Center for Protein Sequences) funded by the Austrian genome initiative GEN-AU (GENome Research in Austria) (Guldener et al., 2006a). The first version of the FGDB was based on the first genome assembly by the Broad Institute (BROAD), whereas the content of FGDB version 3.1 is based on the full manually corrected gene set in the assembly 3 (FG3) by the BROAD (Wong et al., 2011).

Based on the comparison of FG3 and FGDB version 3.1 gene sets, 82% (11,257) genes are the same in terms of exon/intron structures, 2461 genes from FG3 are not present in FGDB version 3.1 or have different structure, whereas 2056 genes from FGDB version 3.1 do not exist in FG3 or have different structure (Wong et al., 2011).

Shortly after, a new gene model set was released as version 3.2. Refined annotation resulted in 13,826 mRNA transcripts. Recently, the genome sequencing and annotation of the PH-1 strain for *F. graminearum* was performed at Rothamsted Research and resulted in RRes version 4.0, the completed genome sequence of the fungus (King et al., 2015). This completed genome sequence is, in effect, a combining of the BROAD Sanger sequencing draft, MIPS version 3.2 gene prediction based on Affymetrix array analyses, whole-genome shotgun re-sequencing to 85-fold coverage, publicly available RNAseq and proteomic datasets, *de novo* assembly, re-annotation of the gene models followed by the manual curation. The completed *F. graminearum* genome is available from Ensembl Fungi (Kersey et al., 2014) and also accessible via the PhytoPath database (Pedro et al., 2016).

Expression data

Another valuable resource for *F. graminearum* information is the Plant Expression Database for Plants and Pathogens (PLEXdb)[20]. This is a public database for large-scale gene expression analysis in plant and plant pathogens. Collaboration between the BROAD and MIPS (Guldener et al., 2006b) led to the design and the validation of the first Affymetrix GeneChip microarray for *Fusarium* genus, based on the first two genes calls for the *F. graminearum* genome. The current FusariumPLEX experiment data sets include the results of 17 experiments in different developmental stage or growth conditions (Dash et al., 2012).

Known pathogenicity-associated genes

Based on information from PHI-base version 3.2 (in December 2009) and other Rothamsted Research resources, there are 44 genes that have been experimentally proven to play crucial

[20] http://plexdb.org/

roles in the pathogenicity of this fungus (Appendix B Tables B-2 to B-5). Based on the information from PHI-base version 3.8, there are 264 *F. graminearum* genes with experimentally proven roles in pathogenicity (Urban et al., 2015b).

2.5.1.2 *Magnaporthe oryzae* genome

Magnaporthe oryzae, the causal agent of rice blast disease, is the most damaging filamentous fungus of rice crops worldwide. The pathogen has become the model organism in the study of pathogenicity and host-pathogen interactions (Ebbole, 2007).

Sequence data

The first complete genome for the rice pathogenic strain *M. oryzae* 70-50 was sequenced using the WGS approach in 2005 and annotated at the Broad Institute. In this section of the thesis, version 6 of the *M. oryzae* genome assembly was available via the Broad Institute. This assembly genome consists of 11, 074 genes and is 41.7 Mb in size.

Known pathogenicity-associated genes

Based on the information from PHI-base version 3.2 and additional BLASTP analysis (see section 4.2.4), there were 60 genes that have been experimentally proven to play crucial roles in the pathogenicity of this fungus (see Appendix B Tables B-8 to B-10). Based on the information from PHI-base version 3.8, there are 486 *M. oryzae* genes with experimentally proven roles in the pathogenicity and 67 effectors genes (Urban et al., 2015b).

2.5.2 Previous work

In this section the earlier work related to the prediction of protein-protein interactome for both *F. graminearum* and *M. oryzae* is described. Furthermore, the previous study on the implementation of the network approach for prediction of candidate genes in *F. graminearum* is also described here.

2.5.2.1 *Fusarium graminearum* protein-protein interaction network

The first *Fusarium graminearum* protein-protein interaction (FPPI) database was published in 2009 (Zhao et al., 2009). *F. graminearum* protein-protein interactions were predicted based on interologs and domain-domain interactions (DDIs). In the interologs (interaction-ortholog) approach (Matthews et al., 2001), a pair of *F. graminearum* proteins is considered as an interacting pair if their corresponding orthologs in other species interact with each other. The FPPI network was built based on interactions of orthologs from seven model organisms listed in the and interactome datasets: the Human Protein Reference Database (HPRD)[21] (Keshava Prasad et al., 2009), the Molecular INTeraction database (MINT)[22] (Licata et al., 2012), the Biological General Repository for Interaction Datasets (BioGRID)[23] (Chatr-aryamontri et al., 2015), IntAct molecular interaction database (IntAct)[24] (Orchard et al., 2014) and the Database of Interacting Proteins (DIP)[25] (Salwinski et al., 2004). The schematic overview for predicting the FPPI network is presented in Figure 2-10 (source: (Zhao et al., 2009)).

Table 2-2 List of species from which *F. graminearum* protein orthologs were identified.

Organism Name	Genome size [Mb]	Predicted PPI
Escherichia coli	4	1138
Saccharomyces cerevisiae	12	35697
Schizosaccharomyces pombe	14	2350
Caenorhabditis elegans	97	723
Drosophila melanogaster	180	1669
Mus musculus	2700	272
Homo sapiens	3 million	4056

Where PPI – protein - protein interaction.

[21] http://www.hprd.org/
[22] http://mint.bio.uniroma2.it/mint/
[23] http://thebiogrid.org/
[24] http://www.ebi.ac.uk/intact/
[25] http://dip.doe-mbi.ucla.edu/

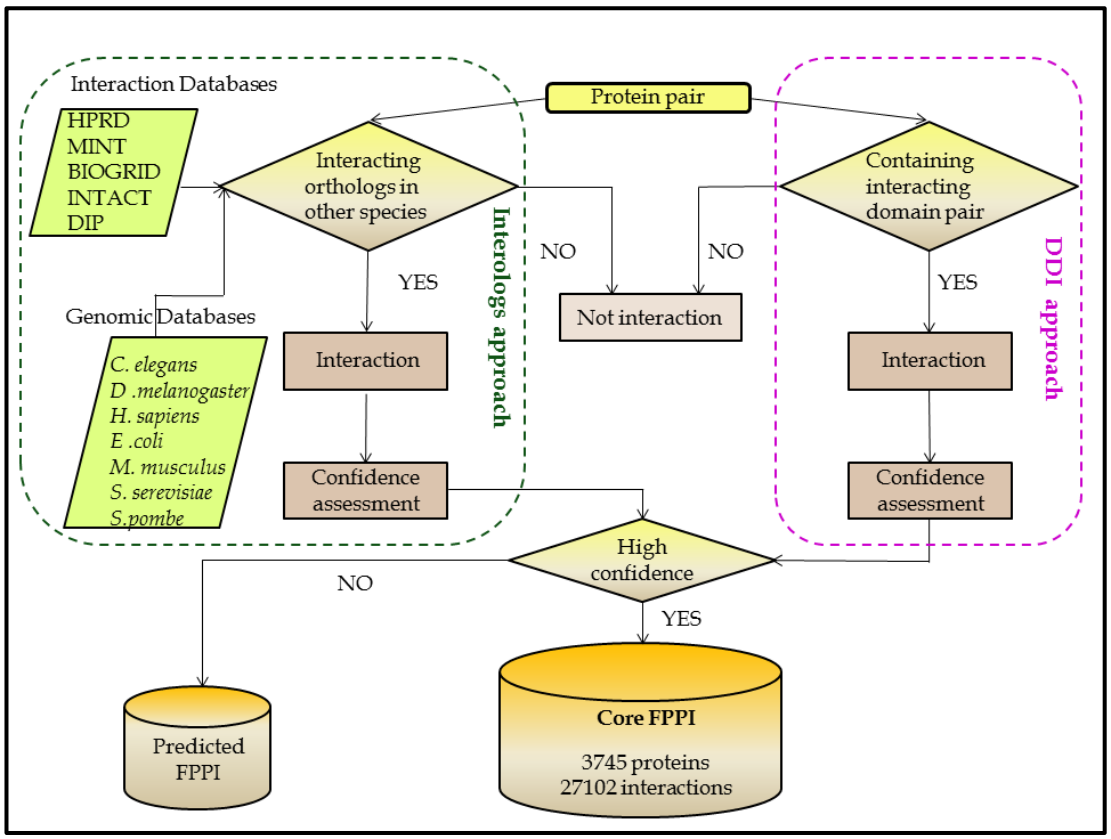


Figure 2-10 Flowchart for prediction of *F. graminearum* protein-protein interactions.
Protein-protein interactions are predicted based on both interologs and domain-domain interaction (DDI) (Zhao et al., 2009).Where FPPI – *F. graminearum* protein-protein interaction database.

Although both *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* are Ascomycete fungi, neither of them are filamentous fungi. Moreover, the genome sizes of these two yeast species each only account for 30% of *F. graminearum* PH-1 strain genome size (36Mb for gene call assembly 3). In addition, none of the seven selected species has a natural pathogenic life cycle.

In the DDI approach, interacting pairs of *F. graminearum* proteins were obtained by mapping the fungus genes to the Pfam-A database (unfortunately the authors did not report the version of Pfam-A database used) to obtain domain signatures for the fungus proteins (or corresponding genes). For every domain found in each gene, the experimenters determined the interacting domains based on the information in iPfam (Finn et al., 2014b) and 3did[26] (3D Interacting

[26] http://3did.irbbarcelona.org/

Domains) (Mosca et al., 2014) databases, as well as by applying a probabilistic model (Equation 2-4) (Zhao et al., 2009):

Equation 2-4

$$P_r(P_{ij} = 1) = 1 - \prod_{d_{mn} \in P_{ij}} (1 - P_r(d_{mn} = 1))$$

Where $(P_{ij} = 1)$ is a probability that protein *i* interacts with protein *j*, $d_{mn} \in P_{ij}$ indicates that domain *m* lies in protein *i* and domain *n* lies in protein *j*, $d_{mn} = 1$ signifies that domain *m* interacts with domain *n.*

Both iPfam and 3did databases are collections of domain–domain interactions from experimentally determined three-dimensional structures of proteins. IPfam is a unique resource integrated into PFAM domain database (Finn et al., 2014a) at the Sanger Institute and 3did is provided by the Structural Bioinformatics and Network Biology Group at the Institute for Research in Biomedicine.

Furthermore, (Zhao et al., 2009) assigned a confidence score for each protein-protein interaction in the FPPI network. In the interologs approach, the confidence score was assigned on the basis that the higher number of interactome datasets and interologs in more species, the higher the confidence score. The confidence score *S* for every pair of interacting proteins was defined by the Equation 2-5 (Zhao et al., 2009).

Equation 2-5

$$S_{ij} = N_{interactome} * N_{species}$$

Where S_{ij} is a confidence score for interacting pair of proteins *i* and *j*, N is a number of interactome data sets and species in which the PPI occurs respectively. In addition, the experimenters set the PPI thresholds for low, medium and high confidence of S_{ij} = 1, 1< S_{ij} <4, and ≥ 4, respectively.

In the case of PPI predicted via DDI, Zhao et al. (2009) calculated PPI in three groups of confidence: low confidence, medium confidence, and high confidence. The classification was made on the assumption that the interacting proteins pair is considered as a medium confidence if each protein in the pair consists of one domain which interacts with the domain of the interacting protein in the pair. Furthermore, the interaction is considered to be of high confidence if the domain in each protein pair covers more than 50% of sequence length of the interacting proteins of medium confidence interactions. Otherwise, the interaction between protein pairs is interpreted as low confidence since the proteins can contain multiple domains or motifs.

2.5.2.2 Prediction of pathogenic genes in *Fusarium graminearum*

Liu and co-workers applied a network approach to predict pathogenic genes for *F. graminearum* (Liu et al., 2010). They used the interactome map for *F. graminearum* (Zhao et al., 2009) described in the previous section. Only high confidence protein interactions from both interologs and DDI were used for the FPPI network (core network). As some of *F. graminearum* genes were experimentally tested for pathogenicity and the information was stored in PHI-base (Baldwin et al., 2006), Liu et al. (2010) mapped 49 *F. graminearum* genes present in PHI-base version 3.1 (see Chapter 4 Table 4-1) into FPPI network in order to predict further pathogenic genes for this economically important pathogen.

In their prediction of pathogenic genes Liu et al. (2010) assumed that a gene in the FPPI network, when connected with at least two *F. graminearum* genes from PHI-base (so-called 'seed' genes), was more likely to be a pathogenic gene. Furthermore, they integrated gene expression data of *F. graminearum* (Guldener et al., 2006b) into the interaction map. Consequently, the obtained sub-network consisted of genes interacting with at least two 'seed' genes *(F. graminearum* genes from PHI-base version 3.1) and the genes had to be differentially expressed *in planta* according to information taken from PLEXdb[27] (Dash et al., 2012).

[27] http://www.plexdb.org/

In addition, they assigned a weight w(x) for each predicted gene taking into account the interaction with seed genes and co-expression data (Equation 2-6):

Equation 2-6

$$w(x) = \sum_{y \in S} PC(x,y) * I(x,y)$$

Where S is a set of seeds genes, $PC(x,y)$ is the correlation coefficient between gene x and gene y, and $I(x,y)$ is an indication function and $I(x,y)$ = 1 if protein x interacts with protein y. Otherwise, $I(x,y)$ = 0. The steps in prediction of *F. graminearum* pathogenic genes are presented as a flowchart in the Figure 2-11 reproduced based on (Liu et al., 2010).

Summarising their study, Liu et al. (2010) mapped in total 20 'seed' genes into the FPPI network and predicted 49 candidate genes for pathogenicity.

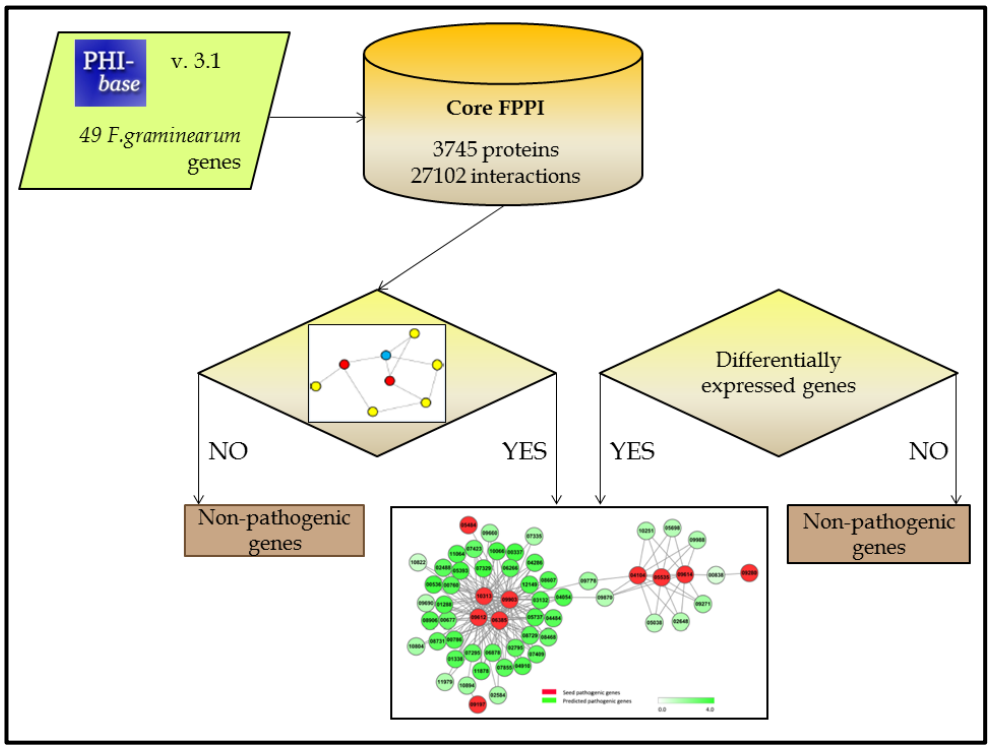


Figure 2-11 Flowchart of steps in the prediction of pathogenic genes in *F. graminearum* (Liu et al., 2010).
Where FPPI – *Fusarium graminearum* protein – protein interaction database.

2.5.2.3 *Magnaporthe oryzae* protein-protein interaction network

The first *Magnaporthe oryzae* protein-protein interaction (MPPI) database was published in 2008 (He et al., 2008). *M. oryzae* protein-protein interactions were predicted using the interologs approach (Matthews et al., 2001) with the assistance of the InParanoid algorithm (Remm et al., 2001). The MPPI network was based on ortholog interactions from five well-studied species listed in Table 2-3 and manually curated interactome datasets: HPRD (Keshava Prasad et al., 2009) and DIP (Salwinski et al., 2004).

In their study, 11,674 interactions within 3,017 *M. oryzae* proteins were predicted. Approximately two-thirds of the interactions were directly predicted from PPI in *Saccharomyces cerevisiae.* The content of Table 2-3 reveals that neither of them is a filamentous fungus. Although *Saccharomyces cerevisiae* represents the Ascomycetes group of fungi, its genome size only accounts for 29% of *M. oryzae* 70-15 strain genome size (41.7Mb). Again, none of the five selected species has a natural pathogenic life cycle or has a filamentous growth habit.

Table 2-3 List of species from which *M. oryzae* protein orthologs were identified.

Organism Name	Genome size [Mb]	Predicted PPI*
Escherichia coli	4	916
Saccharomyces cerevisiae	12	7,904
Caenorhabditis elegans	97	242
Drosophila melanogaster	180	1,080
Homo sapiens	3 million	2,177

Where PPI – protein - protein interaction. *PPI can be inferred from two or more organisms.

In building the MPPI database, interactions were only selected if their orthologs in five different model species (Table 2-3) have at least one experimentally validated interaction.

Chapter 3

Global overview of PHI-base features

3.1 Aim of the study

The aims of the investigation in this chapter are

- General characteristic of the content of PHI-base
- Identifying plant pathogen-specific gene clusters, animal pathogen-specific gene clusters, and those clusters that contain genes associated with both types of pathogens.

Initially, the analyses were performed for PHI-base version 3.2 (December 2010). However, since the content of PHI-base has changed substantially from the time the first analyses were conducted in 2010 (see Table 2-1), it was then interesting to repeat the analyses for the PHI-base version 4.0 to reflect the changes in PHI-base content over the time period covered in this thesis.

3.2 Methods

Since it is known that a member of the same protein family may have similar or even identical functions (Hegyi and Gerstein, 1999), determined by similar protein structure and consequently by similar amino acid sequence, protein families can be described as molecules that share significant sequence similarity. Thus, sequence clustering is a process or algorithm that considers all similarity correlation in a given set of sequences.

Here sequence similarity (Altschul et al., 1997) as a metric for clustering the proteins into families was used. There are different clustering algorithms available such as hierarchical clustering, partitional clustering (K-means, Fuzzy C-means or Quality Threshold (QT) clustering), or spectral clustering. However, a study by Enright et al.(2002) revealed rapid and accurate clustering

method that relies on Markov Cluster (MCL) algorithm[28] (Enright et al., 2002, Van Dongen, 2000) and groups proteins into families based on pre-calculated sequence similarity information. This method was shown to be successful in clustering eukaryotic genomes that consist of proteins with a large number of domains. Based on the above study, the MCL algorithm was chosen to cluster protein sequences present in PHI-base version 3.2 (December 2010). The analysis was then reproduced five years later for PHI-base version 4.0 (September 2015) to highlight the changes in the content of PHI-base in the last half-decade.

3.2.1 Data preparation

Removing redundancy

The FASTA file containing protein sequences (that encodes genes available in PHI-base version 3.2) was directly downloaded from PHI-base website (December 2010). The redundant sequences (117 genes) were removed from the file via a Python script (removeDuplicates.py, see https://github.com/ejsejda/PhD_thesis-Chapter_3) leaving 807 non-redundant protein sequences in the file (FileA). The step was also repeated while analysing the content of PHI-base version 4.0 (September 2015). In PHI-base version 4.0, 95 redundant sequences were removed leaving 3203 non-redundant protein sequences in FileA.

Downloading UniProtKB/TrEMBL sequences

The FASTA file containing UniProtKB/TrEMBL sequences was downloaded from UniProt Knowledgebase (UniProtKB) and saved as FileB.

The Universal Protein Resource (Consortium, 2015) is a centralised repository of protein sequences and annotation data. It consists of four databases: the **UniProtKB** (Protein knowledgebase at UniProt) – central hub for integrated protein information, with accurate consistent and rich annotation; the **UniProt Reference Clusters** (UniRef) – merge closely related sequences based on sequence identity to speed up searches; the **UniProt Archive** (UniParc) – protein sequence archive and **Proteomes** – the set of proteins thought to be expressed by an

[28] http://micans.org/mcl/

organism with completely sequenced genome. Further, **UniProtKB** consists of two sections: **UniProtKB/SwissProt –** manually annotated records with information extracted from literature and curator-evaluated computational analysis; **UniProtKB/TrEMBL** – high quality computationally analyses records that await full manual annotations; quality of sequences depends on the quality of the submissions in the original EMBL-Bank/GenBank/DDBJ entry.

Sequence similarity search

FASTA files: 'FileA' and 'FileB' were merged into one FASTA 'FileC' to obtain a more informative e-value. Hardware acceleration BLASTP on 'FileA' against 'FileC' was performed

3.2.2 Markov Cluster Algorithm (MCL algorithm)

Markov Cluster Algorithm (MCL) - a cluster algorithm for graphs - is a fast cluster algorithm designed specifically for the settings of simple or weighted graphs. Although initially it was used in computational graph clustering, the algorithm was proved to be successful for biological sequences clustering. Sequence similarity relationships among protein sequences set can be represented as a square matrix where each element of the matrix is a number, such e-value from BLAST that represents the similarity for any pair of proteins in the dataset. The matrix then is passed through iterative rounds of matrix multiplication and inflation until there is a small or no net change in the matrix. Then such a matrix is interpreted as a protein family cluster (Enright et al., 2002, Van Dongen, 2000).

The MCL algorithm takes as an input the file in which each line encodes an edge in terms of two labels (such as SequenceA and SequenceB) and a numerical value (such as e-value), all tab separated. The granularity of the output cluster is controlled by only a single parameter –I *inflation.* By default, it is set to 2.0. In order to perform clustering on BLASTP output file, the file had to be converted into input format required by MCL algorithm. Thus, a Perl script (blastpParse_removeTrembl.pl, see: https://github.com/ejsejda/PhD_thesis-Chapter_3) was used to parse the BLASTP output file into a tab delimited file in which each line included a pair of matched protein sequences and e-value calculating for their match. Further, all UniProtTrEMBL

sequences were removed from the file. Then, the newly created file was an input into MCL algorithm for clusters calculations.

Clustering was performed for different inflation parameters in a range from 0.4 to 3.6 with step of 0.4. The inflation parameter set to 1.6 was chosen as the number of generated clusters greatly changed while setting the parameter from 1.2 to 1.6, as well as from 1.6 to 2.0.

Further analysis of the content of PHI-base version 3.2 and PHI-base version 4.0 was performed using a custom developed script in the Python programming language (gettingHostTaxa.py and analysisOfPHIbase_v4.0.py, see https://github.com/ejsejda/PhD_thesis-Chapter_3).

Following the clustering process, clusters with sequences number two or more were graphically represented with aid of NetworkX version 1.1 (displayMCLclusters.py, see https://github.com/ejsejda/PhD_thesis-Chapter_3). NetworkX is a Python package for the generation, manipulation, and study of the structure dynamics and functions of complex networks.[29]

3.3 Overview the content of PHI-base version 3.2

The sequences in PHI-base version 3.2 were clustered based on sequence similarity and represented graphically to give a global overview of the database content in terms of observed phenotype and experimental hosts. Overall 112 clusters with two or more genes were identified (Figure 3-1). Furthermore, within those 112 clusters, 50 clusters were plant pathogen-specific gene clusters with the largest one containing 9 genes; 25 clusters were animal pathogen-specific gene clusters with the largest one consisting of 8 genes; and 32 clusters grouped both animal and plant pathogen genes.

There is one cluster, consisting of 10 genes, where one of those genes (depicted in magenta in Figure 3-1) was experimentally proven to interact with another fungus.

[29] https://networkx.github.io/

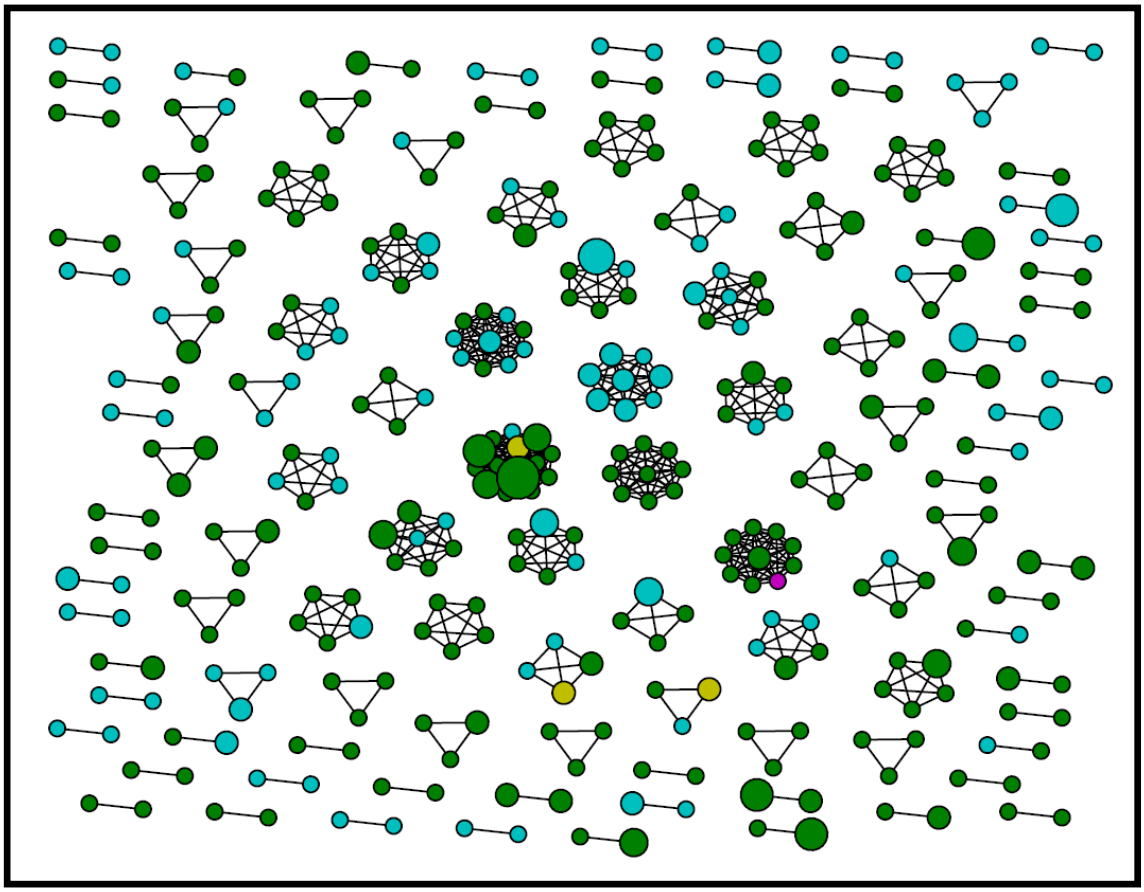


Figure 3-1 Clusters generated in PHI-base version 3.2.
Each node presents a pathogen gene for which interaction with the chosen host was tested. The colour of the node indicates the host the gene interacts with: green, blue, yellow and magenta indicates plant, animal, plant and animal, and other hosts such fungi respectively. The size of each node indicates the number of hosts with which the gene interacts.

3.4 PHI-base version 4.0

Since the data in the PHI-base has changed substantially from the time the analysis in section 3.3 was performed (December 2010, also see Table 2-1), the analysis was repeated with PHI-base version 4.0 (September 2015). General analysis of PHI-base version 4.0, as well as detailed examination of the generated clusters, are presented in this section.

3.4.1 General overview of PHI-base version 4.0

PHI-base version 4.0 contains 3203 unique genes for which protein sequences in FASTA format are available. These genes belong to 265 different pathogenic species including Bacteria, Oomycetes, and Fungi (both Ascomycetes and Basidiomycetes).

By contrast, PHI-base version 3.2 (December 2010) contains only 807 unique genes with available protein sequences in FASTA format. These genes belong to just 75 different pathogenic species including Bacteria, Oomycetes, and Fungi.

General analysis of PHI-base version 4.0 was performed by finding the number of genes experimentally disrupted per particular pathogen, as well as the number of hosts with which a pathogen interacts. The outcome of this analysis is partially presented in Table 3-1 (plant pathogens) and Table 3-2 (animal pathogens). Only species with ten and a higher number of genes, which were experimentally investigated for their pathogenic outcome, are listed in Table 3-1 and Table 3-2. Fungal species such as Fusarium *graminearum*, *Magnaporthe oryzae*, *Ustilago maydis*, *Candida albicans* and *Cryptococcus neoformans* represent well-studied pathogens within PHI-base version 4.0.

A similar trend was observed during the study performed on PHI-base version 3.2 (Appendix A, Table A-1 and Table A-2). However, *F. graminearum* is the most dominant plant pathogen in PHI-base version 4.0 in terms of experimentally verified pathogenicity, virulence, or effector genes number. By contrast to PHI-base version 3.2, we see almost 16 times more *F. graminearum* genes number in PHI-base version 4.0, whereas genes number for *M. oryzae* increased only threefold from version 3.2 to version 4.0 of PHI-base.

Table 3-1 Plant pathogen species in PHI-base version 4.0.
Here species with ten or more experimentally verified pathogenicity, virulence, or effector genes in PHI-base version 4.0 are listed. The third column shows the number of hosts with which a pathogen (column one) interacts. The fourth column lists the number of experimentally verified genes per pathogen.

Pathogen name	**Taxonomy**	**No of hosts**	**No of genes**
Fusarium graminearum (related*: Gibberella zeae*)	Fungi: Ascomycota	14	1033
Magnaporthe oryzae	Fungi: Ascomycota	8	495
Ustilago maydis	Fungi: Basidiomycota	4	251
Botrytis cinerea	Fungi: Ascomycota	20	111
Fusarium oxysporum	Fungi: Ascomycota	15	110
Pseudomonas syringae	Bacteria	8	106
Ralstonia solanacearum	Bacteria	7	92
Hyaloperonospora arabidopsidis	Oomycetes	5	69
Xanthomonas oryzae	Bacteria	2	65
Pseudomonas savastanoi	Bacteria	1	52
Xanthomonas campestris	Bacteria	5	52
Parastagonospora nodorum	Fungi: Ascomycota	2	49
Zymoseptoria tritici (related*: Mycosphaerella graminicola*)	Fungi: Ascomycota	2	46
Erwinia amylovora	Bacteria	5	38
Cochliobolus heterostrophus	Fungi: Ascomycota	2	31
Alternaria alternata	Fungi: Ascomycota	8	29
Phytophthora sojae	Oomycetes	3	29
Alternaria brassicicola	Fungi: Ascomycota	10	28
Phytophthora infestans	Oomycetes	7	26
Fusarium verticillioides	Fungi: Ascomycota	4	25
Burkholderia glumae	Bacteria	3	24
Melampsora larici-populina	Fungi: Basidiomycota	1	24
Claviceps purpurea	Fungi: Ascomycota	1	20
Leptosphaeria maculans	Fungi: Ascomycota	3	20
Xylella fastidiosa	Bacteria	1	18
Sclerotinia sclerotiorum	Fungi: Ascomycota	13	15
Verticillium dahliae	Fungi: Ascomycota	12	15
Cochliobolus carbonum	Fungi: Ascomycota	1	14
Colletotrichum gloeosporioides (related*: Glomerella cingulata*)	Fungi: Ascomycota	5	14
Xanthomonas citri	Bacteria	4	14
Blumeria graminis	Fungi: Ascomycota	3	13
Cryphonectria parasitica	Fungi: Ascomycota	2	12
Penicillium expansum	Fungi: Ascomycota	3	12
Cladosporium fulvum	Fungi: Ascomycota	2	11
Colletotrichum graminicola (related: *Glomerella graminicola*)	Fungi: Ascomycota	2	11
Colletotrichum lagenarium	Fungi: Ascomycota	1	11

Table 3-2 Animal pathogen species in PHI-base version 4.0.
Here species with ten or more experimentally verified pathogenicity, virulence, or effector genes in PHI-base version 4.0 are listed. The third column shows the number of hosts with which a pathogen (column one) interacts. The fourth column lists the number of experimentally verified genes per pathogen. *Model organism.

Pathogen name	Taxonomy	No of hosts	No of genes
Candida albicans	Fungi: Ascomycota	10	240
Salmonella enterica	Bacteria	7	190
Cryptococcus neoformans	Fungi: Basidiomycota	6	120
Aspergillus fumigatus	Fungi: Ascomycota	3	97
Mycobacterium tuberculosis	Bacteria	2	62
Pseudomonas aeruginosa	Bacteria	13	58
Streptococcus pneumoniae	Bacteria	1	57
Staphylococcus aureus	Bacteria	5	55
Riemerella anatipestifer	Bacteria	1	51
Burkholderia pseudomallei	Bacteria	2	45
Escherichia coli	Bacteria	9	44
Vibrio cholerae	Bacteria	3	39
Streptococcus pyogenes	Bacteria	3	25
Beauveria bassiana	Fungi: Ascomycota	5	23
Francisella tularensis	Bacteria	1	23
Yersinia pestis	Bacteria	2	23
Streptococcus suis	Bacteria	5	22
Candida glabrata	Fungi: Ascomycota	2	20
Listeria monocytogenes	Bacteria	2	20
Enterococcus faecium	Bacteria	2	18
Burkholderia cenocepacia	Bacteria	6	17
Enterococcus faecalis	Bacteria	3	16
Klebsiella pneumoniae	Bacteria	2	13
Saccharomyces cerevisiae *	Fungi: Ascomycota	2	13
Vibrio vulnificus	Bacteria	5	13
Yersinia pseudotuberculosis	Bacteria	1	13
Clostridium difficile	Bacteria	3	12
Toxoplasma gondii	Other Eukaryota	2	12
Helicobacter pylori	Bacteria	2	10

3.4.2 Clustering the genes in PHI-base version 4.0

In total 2784 clusters were generated in PHI-base version 4.0, where 361 clusters have two or more genes (Figure 3-2). Furthermore, within those 361 clusters, 243 clusters are plant pathogen-specific gene clusters, 66 clusters are animal pathogen-specific gene clusters and 43 clusters where both animal and plant pathogens were grouped. In addition, a further four clusters were identified with genes experimentally proven to interact with another fungus. Moreover, in this analysis clusters with lethal genes (white nodes) and chemical targets were found.

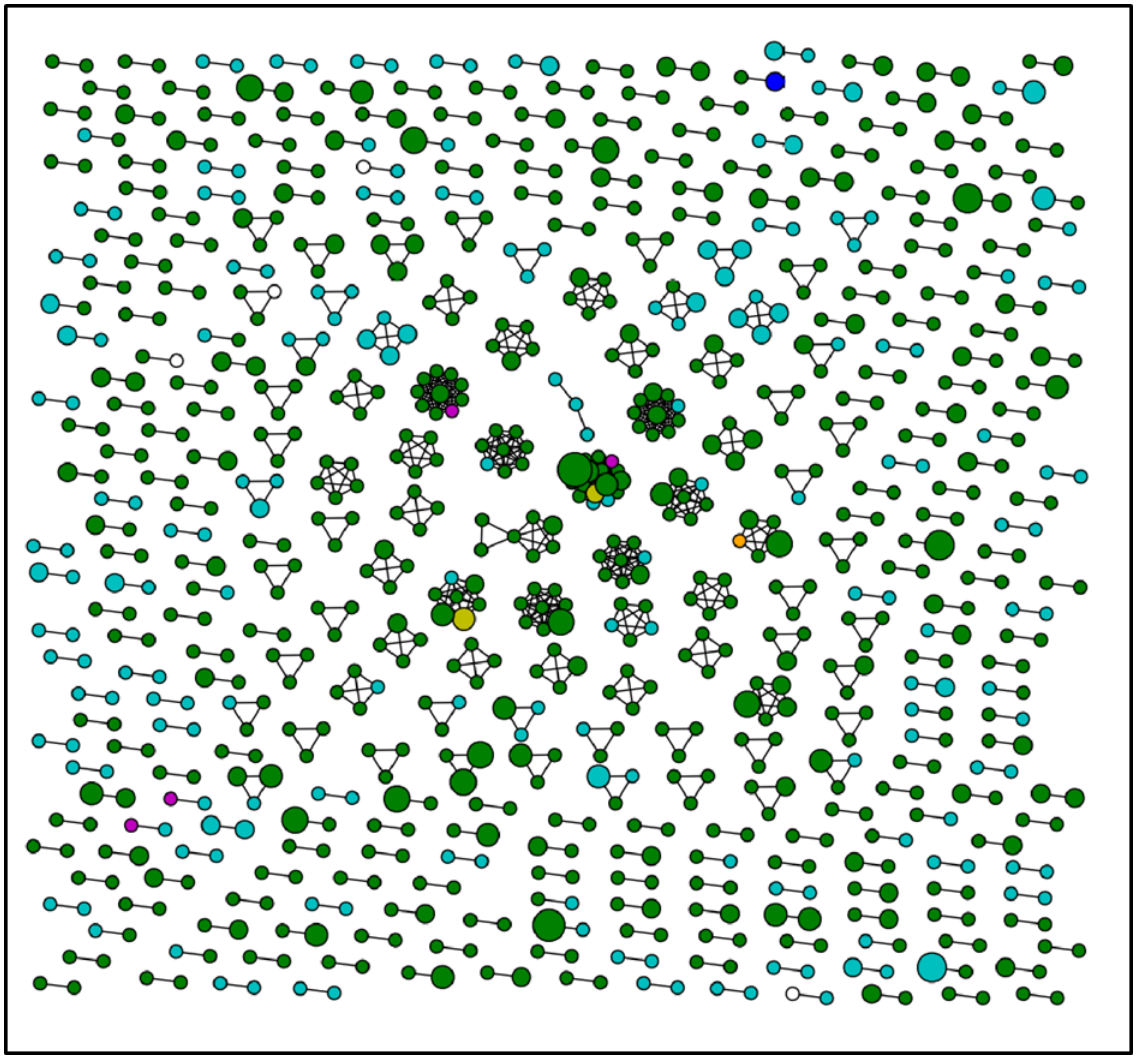


Figure 3-2 Clusters generated in PHI-base version 4.0.
Each node presents a pathogen gene for which interaction with the chosen host (or target protein) was tested. The colour of the node indicates the host the gene interacts with: green, cyan, yellow, blue and magenta indicates plant, animal, plant and animal, other host and animal, and other hosts such fungi respectively. White and orange nodes indicate lethal gene and chemical target respectively. The size of each node indicates the number of hosts with which the gene interacts.

3.4.2.1 Detailed analysis of the largest clusters

The detailed content of three of the largest clusters depicted in Figure 3-2 was further analysed and the results are presented in Figure 3-3 to Figure 3-8, as well as in Table 3-3 to Table 3-5.

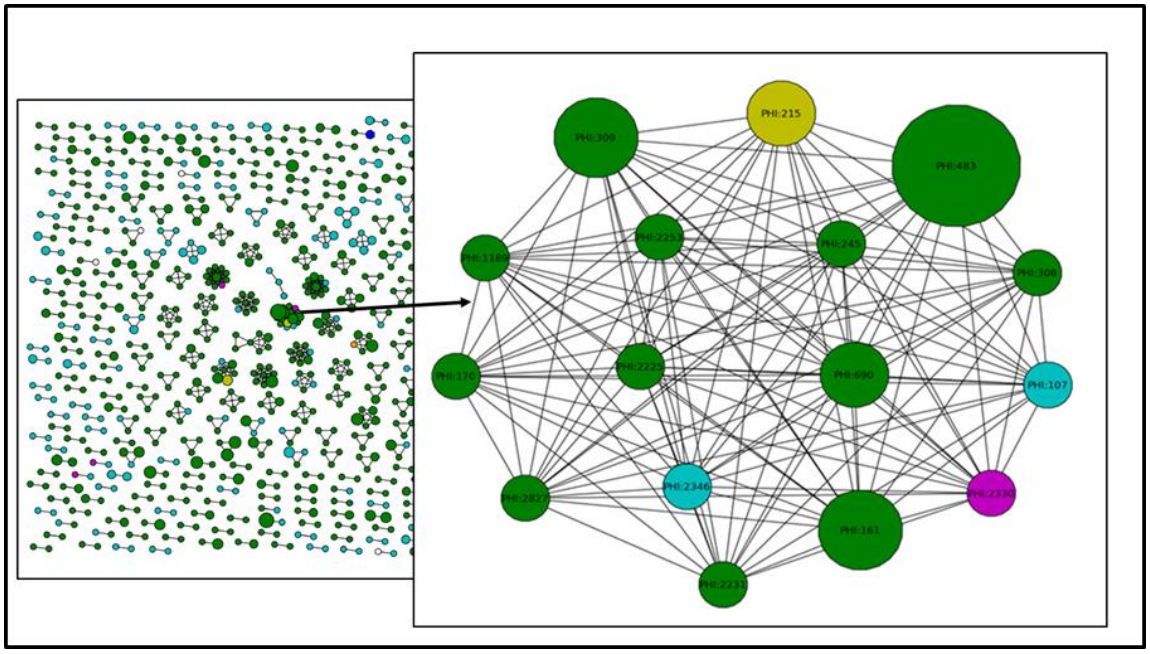


Figure 3-3 Overview of the largest cluster.
Each node presents a pathogen gene for which interaction with the chosen host was tested. The colour of the node indicates the host the gene interacts with: green, cyan, yellow and magenta indicates plant, animal, plant and animal, and other hosts such fungi respectively. The size of each node indicates the number of hosts with which the gene interacts.

As seen from Figure 3-3, the largest cluster is composed of genes from both plant and animal pathogens. Furthermore, within this cluster one fungal pathogen is identified (depicted in magenta in Figure 3-3). Although slight diversity in the observed phenotype can be noticed in the largest cluster (Table 3-3), the disruption to the majority of the genes within this cluster leads to loss or reduced pathogenic activity towards the experimental host.

Table 3-3 Detailed content of the largest cluster.

PHI-base Id	Gene	Gene function	Pathogen name	No of hosts	Phenotype observed
PHI:107	CEK1	MAP kinese	*Candida albicans*	1	reduced virulence (1)
PHI:161	BMP1	MAP kinese	*Botrytis cinerea*	3	loss of pathogenicity (3)
PHI:170	CMK1	MAP kinese	*Colletotrichum lagenarium*	1	loss of pathogenicity (1)
PHI:215	FMK1	MAP kinese	*Fusarium oxysporum*	2	unaffected pathogenicity (1), loss of pathogenicity (1)
PHI:245	CPMK1	MAP kinese	*Claviceps purpurea*	1	loss of pathogenicity (1)
PHI:308	KPP6	MAP kinese	*Ustilago maydis*	1	reduced virulence (1)
PHI:309	MAP1	MAP kinese	*Fusarium graminearum*	2	loss of pathogenicity (2), reduced virulence (1)
PHI:483	VMK1	MAP kinese	*Verticillium dahliae*	7	reduced virulence (6), loss of pathogenicity (1)
PHI:690	PMK1	MAP kinese	*Magnaporthe grisea*	2	loss of pathogenicity (2)
PHI:1189	GMPK1	MAP kinese	*Fusarium graminearum*	1	reduced virulence (1)
PHI:2225	SHO1	High osmolarity signalling protein	*Ustiligo maydis*	1	reduced virulence (1), loss of pathogenicity (1)
PHI:2231	ROK1	MAP kinese	*Ustilago maydis*	1	reduced virulence (1)
PHI:2253	FHK1	MAP kinese	*Fusarium oxysporum*	1	reduced virulence (1)
PHI:2330	LfPMK1	MAP kinese	*Lecanicillium_fungicola*	1	unaffected pathogenicity (1)
PHI:2346	BBMPK1	MAP kinese	*Beauveria bassiana*	1	loss of pathogenicity (1)
PHI:2827	FGB1	G-protein subunit	*Fusarium oxysporum*	1	loss of pathogenicity (1)

The figure in brackets shows the number of phenotypes observed per each gene.

Moreover, investigation of proteins function encoded by the genes within the largest cluster showed that in majority MAP kinases genes are residing within this genes cluster. Further analysis of the first largest cluster reveals a very good conservation within the aligned protein sequences (Figure 3-4). This is especially noticeable throughout the whole region presented in Figure 3-4. The alignment depicted in Figure 3-4 was built with MUSCLE algorithm (Edgar, 2004) and the figure was created with aid of JalView software (Waterhouse et al., 2009).

The second largest cluster, depicted in Figure 3-5, is mainly comprised of genes from fungal plant pathogens. The exception here is a *C. neofomans gene* (STE12a), which is responsible for virulence towards the animal host, namely *Mus musculus* (house mouse). Quite significant homogeneity in observed phenotype is visible here (Table 3-4). All genes, except one (PHI:1387), within this cluster are responsible for pathogenic activities towards the examined host and disruption to these genes leads to either total or reduced loss of pathogenic activity towards the host.

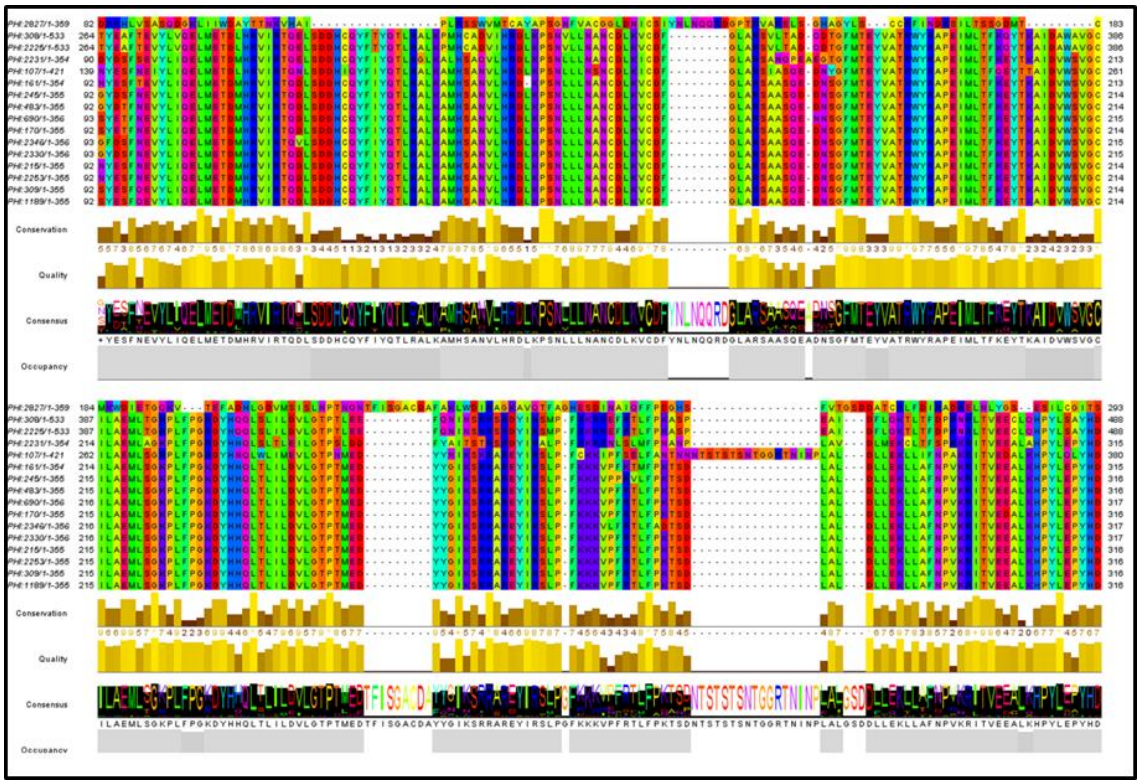


Figure 3-4 Fragment of multiple alignment of the sequences from the largest cluster.

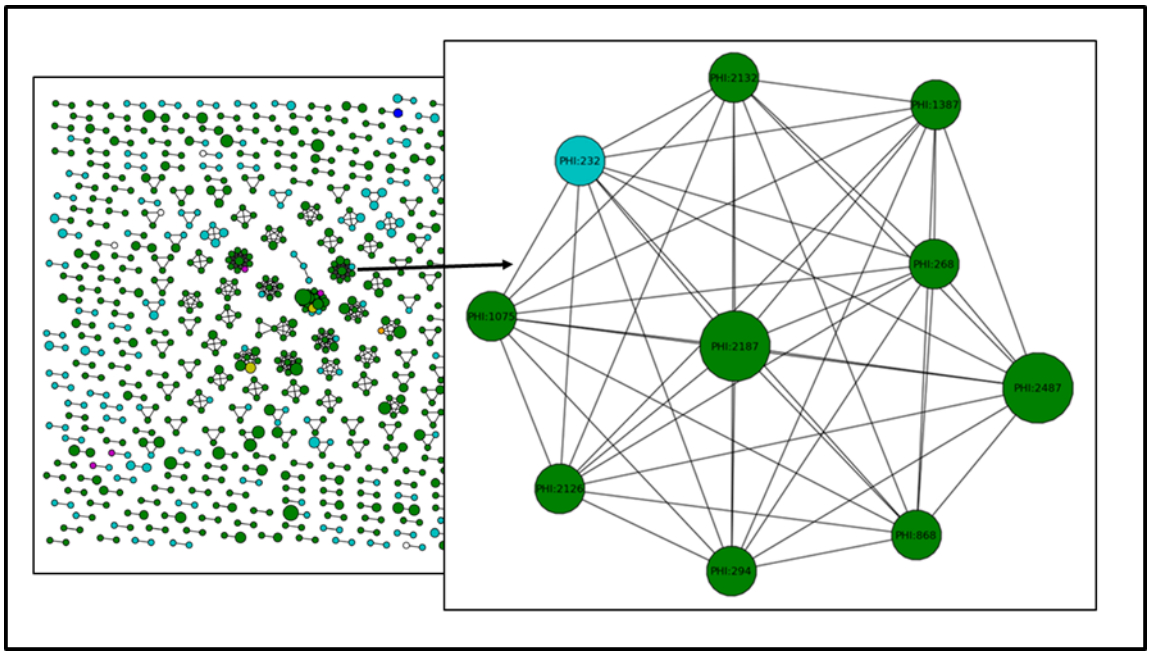


Figure 3-5 Detailed analysis of the second largest cluster.
Each node presents a pathogen gene for which interaction with the chosen host was tested. The colour of the node indicates the host the gene interacts with: green and cyan indicate plant and animal hosts respectively. The size of each node indicates the number of hosts with which the gene interacts.

Table 3-4 Detailed content of the second largest cluster.

PHI-base Id	Gene Id	Gene function	Pathogen name	No of hosts	Phenotype observed
PHI:232	STE12a	transcription factor	*Cryptococcus neoformans*	1	reduced virulence (1)
PHI:268	MST12	transcription factor	*Magnaporthe oryzae*	1	loss of pathogenicity (1)
PHI:294	CST1	transcription factor	*Colletotrichum lagenarium*	1	loss of pathogenicity (1)
PHI:868	CLSTE12	transcription factor	*Colletotrichum lindemuthianum*	1	loss of pathogenicity (1)
PHI:1075	STE12	transcription factor	*Zymoseptoria tritici*	1	reduced virulence (1)
PHI:1387	GzBrlA	transcription factor	*Fusarium_graminearum*	1	unaffected pathogenicity (1)
PHI:2126	MGSTE12p	transcription factor	*Zymoseptoria_tritici*	1	reduced virulence (1)
PHI:2132	MoHOX8	transcription factor	*Magnaporthe oryzae*	1	loss of pathogenicity (1)
PHI:2187	MoMCM1	transcription factor	*Magnaporthe oryzae*	2	loss of pathogenicity (2)
PHI:2487	STE12	transcription factor	*Botrytis_cinerea*	2	reduced virulence (2)

The figure in brackets shows the number of phenotypes observed per each gene.

Further examination of Table 3-4 revealed that all genes comprising the second largest cluster are transcription factor proteins. Moreover, significant conservation regions within the aligned protein sequences of the second largest cluster are observed (Figure 3-6). This is especially noticeable throughout the alignment fragment depicted in Figure 3-6. The sequence alignment presented in the figure was built with MUSCLE algorithm (Edgar, 2004) and the figure was created with aid of JalView software (Waterhouse et al., 2009).

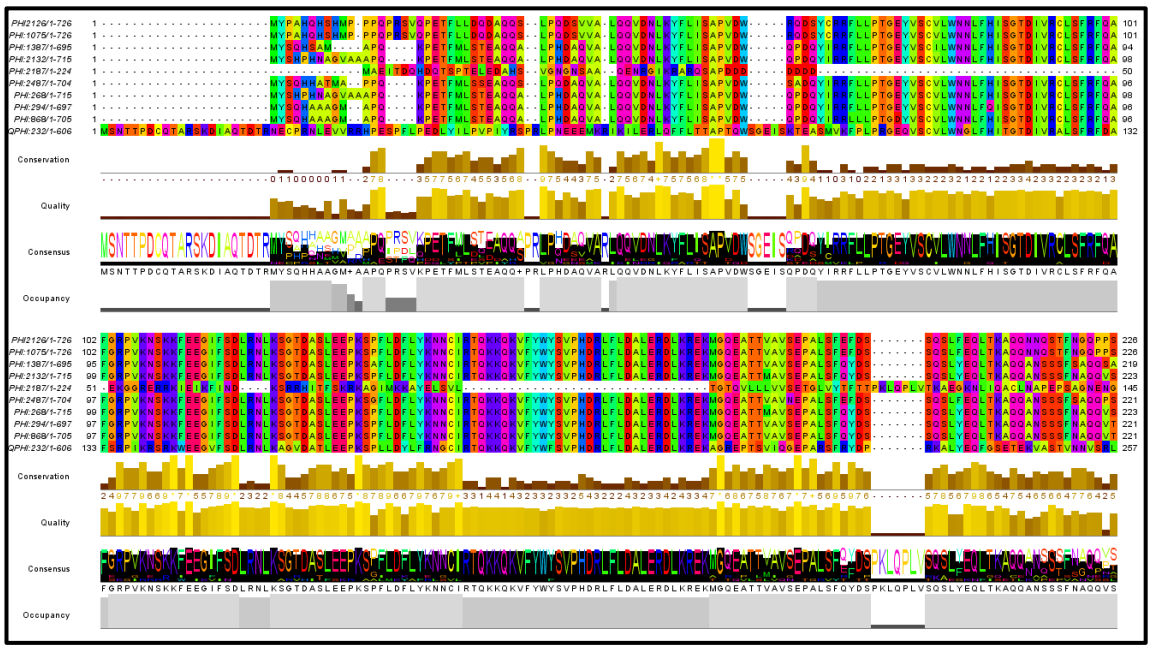


Figure 3-6 Fragment of multiple alignment of the sequences from the second largest cluster.

The third largest cluster is mainly composed of genes that belong to plant pathogens (Figure 3-7). However, one of the genes, depicted in magenta in Figure 3-7, comes from the fungal pathogen, namely *Trichoderma virens,* that infects other fungi.

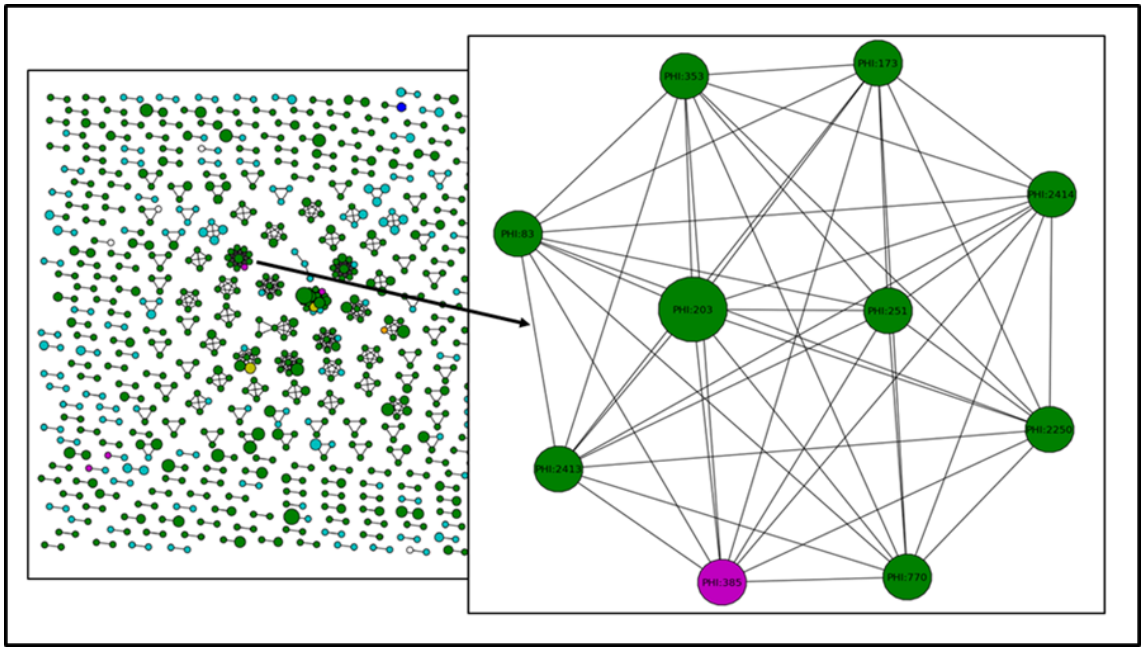


Figure 3-7 Detailed analysis of the third largest cluster.
Each node presents a pathogen gene for which interaction with the chosen host was tested. The colour of the node indicates the host the gene interacts with: green and magenta indicate plant and other hosts such fungi respectively. The size of each node indicates the number of hosts with which the gene interacts.

Furthermore, as per the previous largest cluster examples, fairly substantial homogeneity in observed phenotype is evident here (Table 3-5). All genes, except two: PHI:770 and PHI:2414, within this cluster are responsible for pathogenic activities towards the examined host and disruption to these genes leads to either total or reduced loss of pathogenic activity towards the host.

Further examination of Table 3-5 reveals that all genes comprising the third largest cluster are G proteins known as guanine nucleotide binding proteins. Additionally, regions of considerable conservation within the aligned protein sequences of the third largest cluster is observed here. This is especially noticeable throughout the alignment fragments depicted in Figure 3-8. The sequence alignment presented in Figure 3-8 was built with MUSCLE algorithm (Edgar, 2004) and Figure 3-8 was created with aid of JalView software (Waterhouse et al., 2009).

Table 3-5 Detailed content of the third largest cluster.

PHI-base Id	Gene Id	Gene function	Pathogen name	No of hosts	Phenotype observed
PHI:83	MAGB	G-protein	*Magnaporthe oryzae*	1	reduced virulence (1)
PHI:173	CTG1	G-protein	*Colletotrichum trifolii*	1	loss of pathogenicity (1)
PHI:203	BCG1	G-protein	*Botrytis cinerea*	2	reduced virulence (2)
PHI:251	FGA1	G-protein	*Fusarium oxysporum*	1	reduced virulence (1)
PHI:353	GNA1	G-protein	*Parastagonospora nodorum*	1	reduced virulence (1)
PHI:385	tgaA	G-protein	*Trichoderma virens*	1	reduced virulence (1)
PHI:770	CPG2	G-protein	*Cryphonectria parasitica*	1	unaffected pathogenicity (1)
PHI:2250	GNA1	G-protein	*Parastagonospora nodorum*	1	reduced virulence (1)
PHI:2413	CGG1	G-protein	*Colletotrichum graminicola*	1	reduced virulence (1)
PHI:2414	AaGa1	G-protein	*Alternaria alternata*	1	unaffected pathogenicity (1)

The figure in brackets shows the number of phenotypes observed per each gene.

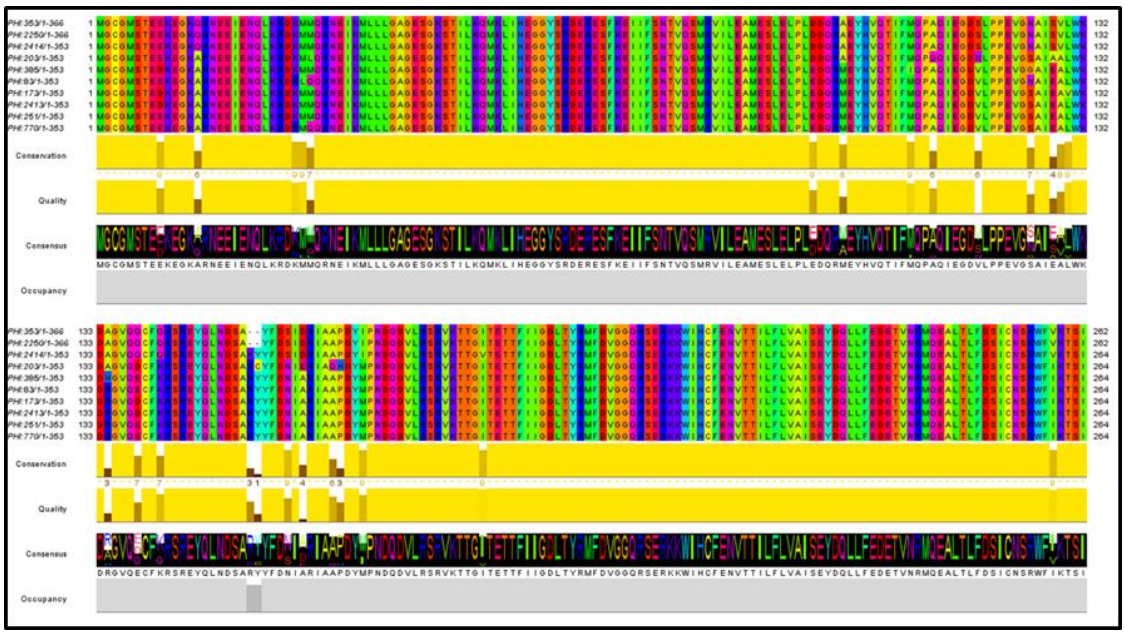


Figure 3-8 Fragment of multiple alignment of the sequences from the third largest cluster.

In conclusion, detailed study of the three largest clusters validates the use of the MCL algorithm for biological sequences clustering.

3.5 Summary

The analysis in this chapter reflected on changes of the database content and size throughout the years of the study reported in this thesis. Here, the comparison of the content of the PHI-base version 3.2 to the content of the PHI-base version 4.0 was performed. The main emphasis was then placed on the detailed analysis of the content of PHI-base version 4.0.

In both versions of PHI-base, clusters that contain genes associated with plant, animal, and both types of pathogens were identified. Moreover, while exploring in detail the clusters generated in PHI-base version 4.0, significant homogeneity was noticed among protein sequences comprising examined clusters. The consistency within each cluster was much noticeable in terms of gene function encoded by protein sequences being part of the same cluster. Additionally, the disruption to genes within the same cluster overall resulted in same/similar observed phenotype.

The consistency within each cluster was confirmed based on several features including gene function, phenotypic outcome, or a clade. This finding suggests that there exist common, as well as unique pathogenicity determinants for plant and animal species. In addition, cereal invading filamentous fungal species have a unique gene repertoire which has enabled them to become successful plant pathogens.

Furthermore, while comparing the content PHI-base version 3.2 to the content of PHI-base version 4.0, we can observe a considerable increase in the number of species catalogued within the database, as well as genes per species present in both database versions. Moreover, the study revealed that five fungal species: *Fusarium graminearum*, *Magnaporthe oryzae*, *Ustilago maydis*, *Candida albicans*, and *Cryptococcus neoformans* represent well-studied pathogens within both PHI-base versions. Two of these fungal species, namely *Fusarium graminearum* and *Magnaporthe oryzae* appeared to be the most overrepresented plant fungal pathogens within both PHI-base versions. Thus, both species were chosen for further analysis within next chapters of the thesis.

Chapter 4

Prediction of pathogenicity genes with the aid of an existing Protein-Protein Interaction (PPI) network

Successful identification of PHI-base genes associated with plant, animal, and both pathogens performed in Chapter 3, as well as the availability of predicted protein-protein interactomes for two economically important plant pathogenic species, namely *Fusarium graminearum* (FG) and *Magnaporthe oryzae* (MO) (Zhao et al., 2009, He et al., 2008) together with the previous study (Liu et al., 2010) on the implementation of the network approach for prediction of pathogenic genes in FG (described in Chapter 2) led to the analysis performed and described in this chapter.

Here a PPI network approach together with experimentally predicted pathogenic genes for both FG and MO, available within PHI-base version 3.2, are used to identify candidate pathogenicity genes in these two economically important plant pathogenic species.

4.1 Aim of the study

Inspired by the previous study (Liu et al., 2010), and with an increasing number of experimentally confirmed pathogenicity and virulence genes now known for a number of plant pathogenic fungi, further development of the PPI network approach to predict candidate genes responsible for pathogenicity was performed. As highlighted in Chapter 3, two well represented in PHI-base plant pathogenic fungi, namely *F. graminearum* and *M. oryzae* were selected for this study.

In contrast to the previous study (Liu et al., 2010), experimentally verified *F. graminearum* genes were carefully dissected into those affecting pathogenicity and those that were shown experimentally not to affect the pathogenicity process. Furthermore, investigation of each FPPI network dataset for both the DDI and Interologs approaches was conducted individually with respect to the different confidence level. For each analysis of the FPPI network two different rules scenarios of predictions were applied. Moreover, network properties such as clustering coefficient, degree centrality, and betweenness were explored.

In addition, characterisation of the *F. graminearum* genome based on functional categories of the genes was performed. Finally, identification of the position of the predicted pathogenic genes within the chromosomes for *F. graminearum* was determined.

Furthermore, with the availability of predicted PPI network for rice blast fungus *M. oryzae* (He *et al., 2008)*, the concept of employing a PPI network approach into prediction of candidate genes responsible for the pathogenicity was utilised for *M. oryzae*.

4.2 Data and methods

4.2.1 The input data

Fusarium graminearum

The predicted *F. graminearum* interactome datasets were provided to us by Dr Xiaoping Liu from Institute of Systems Biology at Shanghai University, China. Furthermore, based on the information in PHI-base version 3.2 and additional Rothamsted Research resources, *F. graminearum* genes known in terms of the proven role on the pathogenicity process were used in the prediction. These genes were dissected into a further five groups that are listed in the Appendix B (Tables B-2 to B-6). In Table 4-1 a general overview is presented, whereas the additional five tables in Appendix B (Tables B-2 to B-6) show the detailed characterisation of those genes with the FGSG number (the gene accession number from the Broad Institute annotation) and functional characterisation, both provided by Dr Martin Urban from Rothamsted Research.

Additionally, in Table 4-1 based on the information available in PHI-base version 3.1, dissection according to the observed phenotype of *F. graminearum* genes present in PHI-base version 3.1 was performed.

Table 4-1 Summary of *F. graminearum* genes with assigned phenotype for PHI-base versions 3.1 and 3.2

Observed Phenotype	PHI-base ver. 3.2	PHI-base ver. 3.1 –genes used by (Liu et al., 2010)
	No of Genes	
Loss of Pathogenicity	3	2
Reduced Virulence	35	21
Reduced Virulence / Unaffected (host-dependent)	3	2
Reduced Virulence / Loss of pathogenicity (host-dependent)	0	2
Increased Virulence	3	0
Unaffected	34	22
TOTAL	**78**	**49**

Magnaporthe oryzae

The predicted *M. oryzae* interactome dataset was downloaded from the link in supplementary materials of study by He et al. (2008). Furthermore, based on the information in the PHI-base version 3.2, the genes for *M. oryzae* strain 70-15 were selected. However, not all of the PHI-base genes for the 70-15 strain of *M. oryzae* were used in this analysis. This is because the genes present in the PHI-base version 3.2 are based on the genome assembly 5 of the *M. oryzae* 70-15 strain. In the newer *M. oryzae* 70-15 strain genome assembly 6, some of the genes that were present in the earlier genome assembly were removed. Moreover, some of the genes of a 70-15 strain of *M. oryzae* which are present in the 3.2 version of PHI-base did not have a corresponding MGG number (the gene accession number from Broad Institute annotation) assigned.

In order to assign these missing MGG accession numbers to selected *M. oryzae* genes in the PHI-base version 3.2, BLASTP (Basic Local Alignment Search Tool for Proteins) a similarity search was performed on all genes from *M. oryzae* 70-15 strain, present in PHI-base version 3.2, against genome assembly 6 of *M. oryzae* 70-15 strain (see section 4.2.4). Tables B-8 to B-11 in the Appendix B list PHI-base version 3.2 *M. oryzae* 70-15 strain genes that were used in the prediction of candidate pathogenicity genes in *M. oryzae* together with MGG accession number. All genes listed in the Appendix B Tables B-8, B-9 and B-10 are known in terms of their proven role on the pathogenicity. Table 4-2 presents the general overview of this gene set.

Table 4-2 Summary of *M. oryzae* strain 70-15 genes present in PHI-base version 3.2.

Observed phenotype	PHI-base version 3.2 used in the prediction
	No of genes
Loss of pathogenicity	6
Reduced virulence	50
Loss of pathogenicity / reduced virulence	4
Unaffected pathogenicity	2
TOTAL	**62**

The genes were used in the prediction of candidate genes for *M. oryzae*.

4.2.2 Analysis and calculation of network properties

Constructing, analysis and calculation of networks properties were performed with the aid of python package NetworkX version 1.1. The visualisation of the *F. graminearum* predicted network was performed via Ondex[30] software (Köhler et al., 2006).

NetworkX is a Python language package for the creation, manipulation, and the analysis of the structure, dynamics, as well as functions of complex networks.

Ondex is a data integration platform that enables linkage of the data from diverse biological data sets. It is also a software for visualisation and analysis of the graphs (Köhler et al., 2006, Lysenko et al., 2009).

4.2.3 Statistical analysis

Statistical analysis of the generated results was performed with the aid of R software version 2.13.1[31] R is a free software environment for statistical computing and graphics. Non-parametric Wilcoxon Rank Sum test with continuity correction and two-sample Kolmogorov-Smirnov statistical test were used to compare network properties distributions.

[30] http://www.ondex.org
[31] http://www.r-project.org/

4.2.4 Mapping PHI-base IDs for *M. oryzae* strain 70-15 genes to MGG IDs

BLASTP similarity search, with p-value threshold set to 10^{-6}, was performed on all *M. oryzae* 70-15 strain genes from PHI-base version 3.2 against assembly 6 genome of *M. oryzae* 70-15 strain.

4.3 Prediction and characterisation of *Fusarium graminearum* candidate genes

4.3.1 Functional characterisation of *Fusarium graminearum* genes using FunCat ontology

In this analysis, functional categorisation of the *F. graminearum* predicted proteome was determined based on FunCat categories (Ruepp et al., 2004) (Figure 4-1). The use of the hierarchical FunCat system was preferred over GO in this analysis due to its ease of perception and because the FunCat functional categorisation was used for the annotation of the *F. graminearum* BROAD genome assembly by MIPS. The functional categorisation in this analysis was based on general level 1 of the FunCat classification. The FunCat content at the time of analysing the data in 2010 consisted of 28 main functional categories listed in Appendix B Table B-7. The FunCat categories for all FPPI network datasets were determined and the results are presented in Table 4-3.

The data in Figure 4-1 and Table 4-3 indicates that some of the functional categories such as metabolism (01), information pathway (16), transport (20) or localisation (70) were very well represented in the *F. graminearum* proteome. Moreover, the fact that only a low number of genes functionally characterised as transposable elements were observed indicates that *F. graminearum* does not have many potentially mobile genetic elements within its genome comparing to *F. oxysporum* f. sp *lycopersici* (Cuomo et al., 2007, Ma et al., 2010). It was also interesting to note (Figure 4-1) that there were not any genes with a predicted protein storage function within the *F. graminearum*. This is because in fungi, including *F. graminearum,* storage nutrients for example compatible solutes such as trehalose, mannitol, GABA, lipids and not proteins fulfil this important cellular function. Using the data in Table 4-3, the *F. graminearum* FunCat categories were compared throughout the different FPPI network datasets. The number

of proteins within a particular FunCat category tends to increase in both DDI and interologs from high confidence dataset towards low confidence dataset. The exceptions to this trend are highlighted in yellow and include FunCat categories not commonly present within the predicted proteome.

Table 4-3 MIPS functional categorisation (FunCat) for *F. graminearum* proteome and all FPPI network datasets based on level-1 of FunCat classification.

MIPS FunCat No	FG proteome protein number	Proteins from different data sets in FPPI network							
		DDI			I				DDI + I
		H	M	L	H+M+L	H	M	L	H
1	2501	771	851	1447	874	218	666	828	940
2	444	135	127	276	202	51	160	185	172
4	1	1	0	1	1	1	1	1	1
10	634	90	154	354	471	282	401	464	324
11	702	72	165	363	485	298	429	476	328
12	371	65	95	199	300	128	280	299	170
14	969	213	321	570	630	329	524	613	456
16	1905	471	569	1145	1157	574	978	1119	907
18	239	50	89	161	177	115	152	176	138
20	1507	289	336	652	536	219	422	515	447
30	350	86	133	217	211	85	156	208	142
32	902	254	285	500	317	147	249	308	356
34	589	102	132	261	228	79	170	218	154
36	150	50	59	97	84	51	70	80	85
38	7	2	2	1	2	0	2	2	2
40	224	60	79	122	156	80	124	153	111
41	135	36	43	74	82	28	54	79	55
42	641	164	203	344	387	195	331	378	297
43	332	86	100	169	191	89	140	185	144
45	42	10	15	25	24	11	18	21	16
47	57	7	13	28	29	11	20	27	16
70	2684	593	681	1360	1342	598	1109	1301	1050
73	18	5	6	9	6	1	3	6	5
75	34	8	11	14	16	7	11	16	13
77	86	36	33	55	35	7	24	33	42
98	403	75	105	182	151	56	111	139	121
Sum	15927	3731	4607	8626	8094	3660	6605	7830	6492

Where MIPS – Munich Information Center for Protein Sequences, MIPS FunCat No – please refer to Table B-7 in Appendix B, FG – *Fusarium graminearum*, FPPI – *Fusarium graminearum* protein-protein interaction; H – high confidence, M – middle confidence, L – low confidence; I – interologs approach, DDI – domain-domain interaction approach. Highlighted in yellow proteins number from FunCat categories not commonly present within the predicted *F. graminearum* proteome.

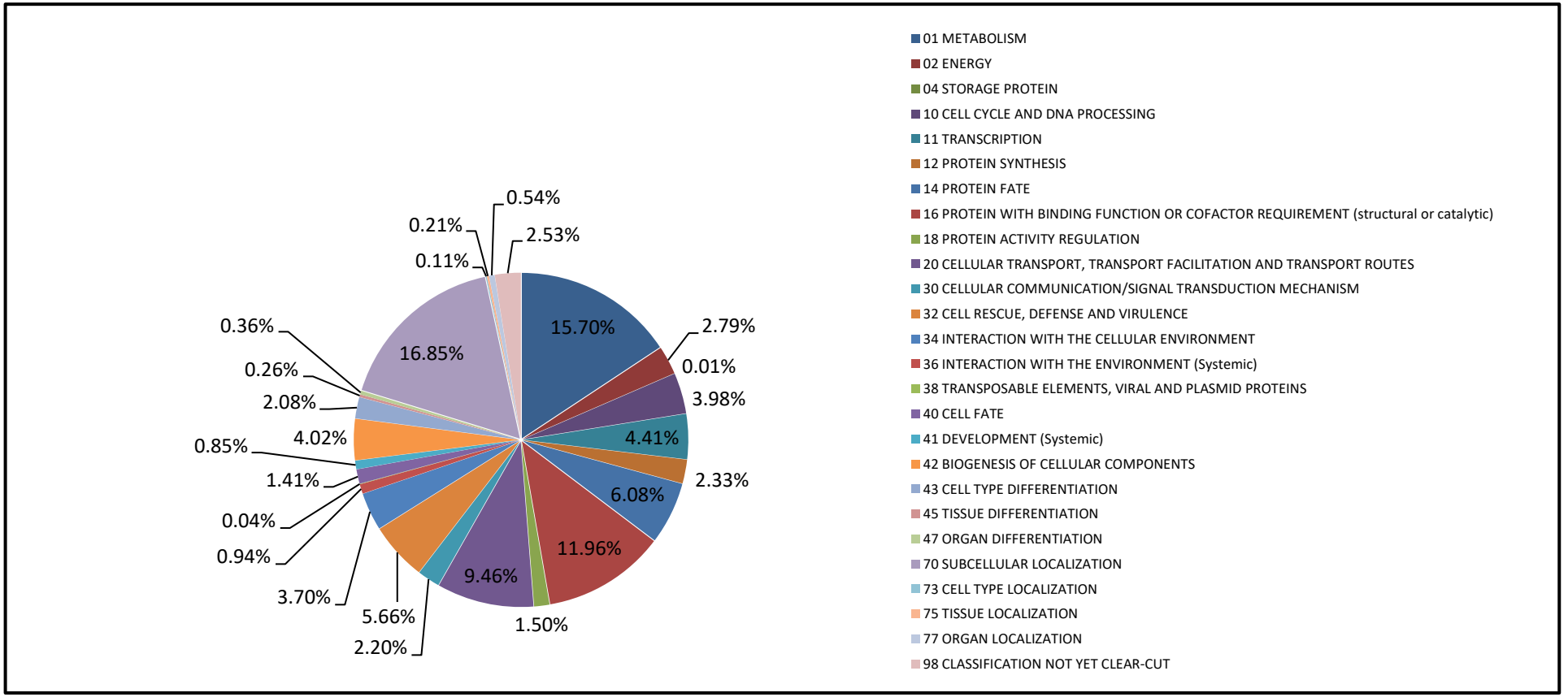


Figure 4-1 MIPS functional categorisation (FunCat) for *F. graminearum* genome based on level 1 of FunCat classification.

In addition to the results presented in Figure 4-1 and Table 4-3, the number of functional categories per predicted protein for the *F. graminearum* genome was calculated. These outcomes are presented in Figure 4-2.

Inspection of Figure 4-2 reveals that the majority of the proteins in *F. graminearum* genome have two predicted FunCat categories, whereas one protein FGSG_09612 has 16 MIPS functional categories assigned. These are 01, 02, 10, 11, 14, 16, 18, 30, 32, 34, 36, 41, 43, 45, 47, and 70 (please refer to Appendix B Table B-7 for the description of FunCat categories). This protein, the *Hog1* MAP kinase, is responsible for the adaptation of *F. graminearum* to hyperosmotic stress. The study by Ramamoorthy et al. (2007), revealed that the *Hog1* MAP kinase signalling cascade is necessary for protection of the pathogen from antimicrobial protein produced by the host plant. Thus, the *HOG1* MAP kinase contributes to pathogenicity.

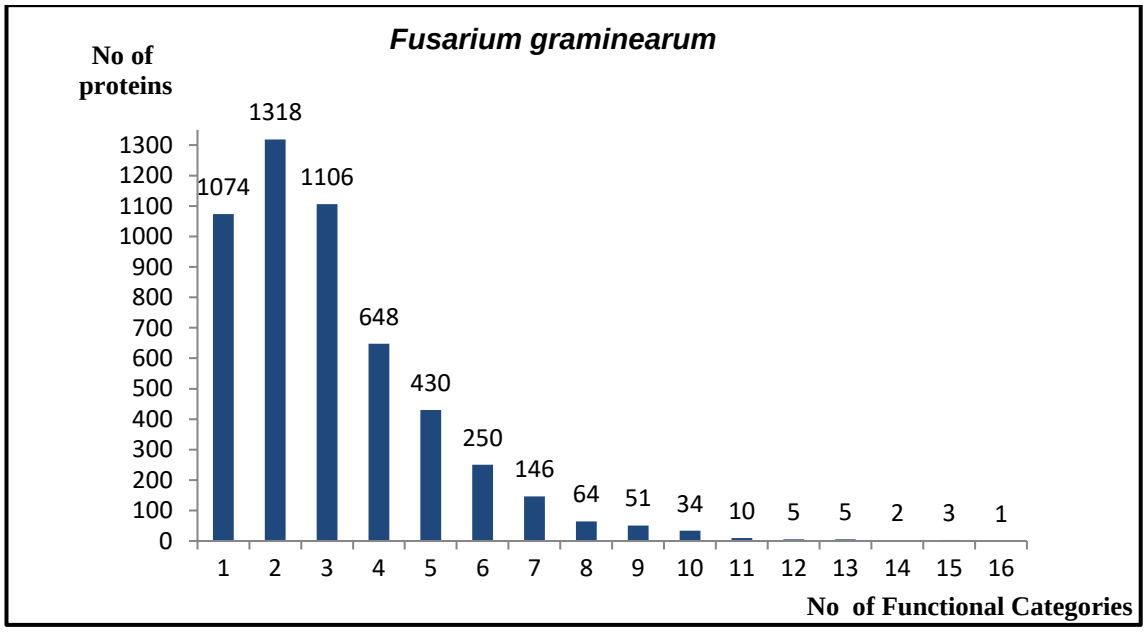


Figure 4-2 Distribution of the number of MIPS FunCat categories for proteins in *F. graminearum* genome.

4.3.2 Using the network for prediction

In this study the interactome datasets for *F. graminearum* (FG) provided by Dr Xiaoping Liu from Institute of Systems Biology at Shanghai University, China were used. Each dataset was examined separately. In order to find candidate genes for the prediction, two scenarios (Figure 4-3) were applied in each examination.

The first scenario used by Liu et al. (2010) (Figure 4-3 A) assumes that the unknown gene is the candidate gene if it is connected to at least two 'seed' genes, while in the second more stringent scenario (Figure 4-3 B) the unknown gene is the candidate gene if it is connected to three 'seed' genes. A 'seed' gene is defined as a FG gene that has been experimentally tested and showed to have a role in pathogenic properties or was shown not to have a role in pathogenicity. All 'seed' genes were dissected according to their phenotype (see Table 4-1 and Appendix B Tables B-2 to B-6) and are decorated in red colour in the depictions.

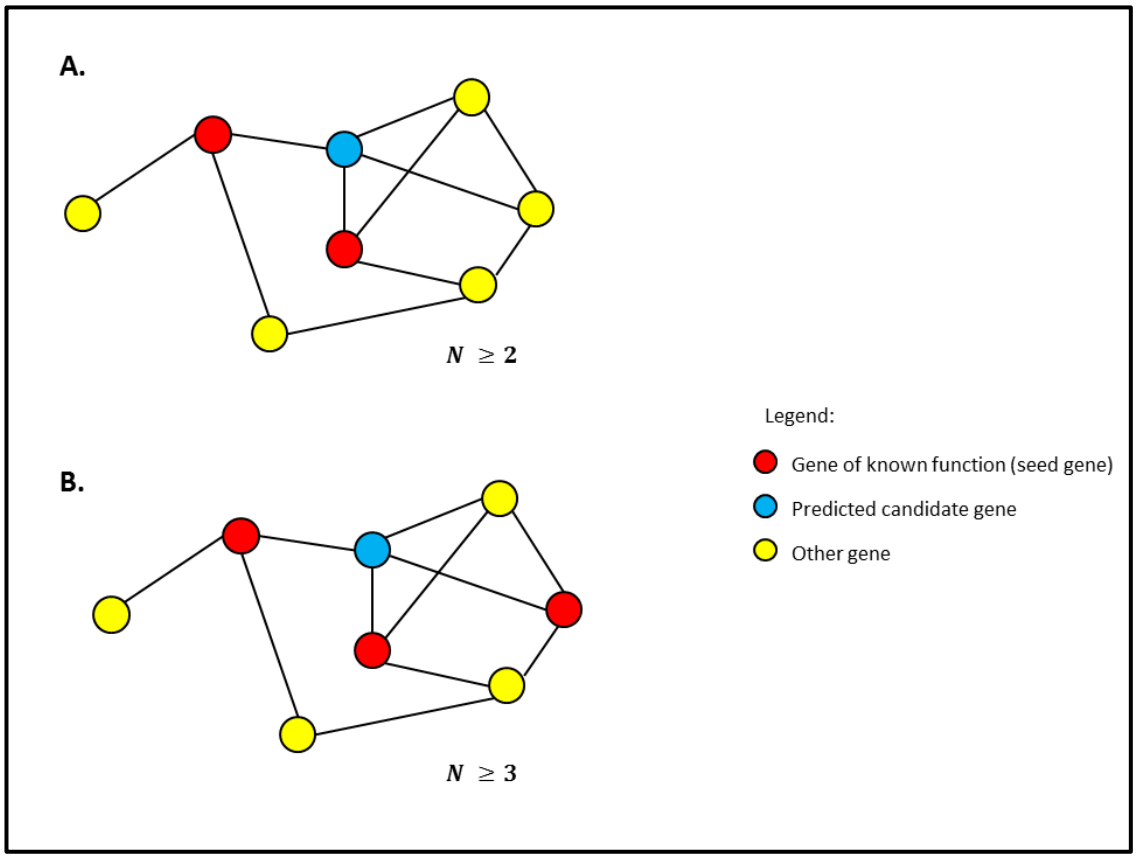


Figure 4-3 Two different scenarios for assigning pathogenic genes in the FPPI network.
Where 'seed' genes are depicted in red and predicted candidate genes are depicted in blue. N indicates the minimum number of 'seed' genes to be connected to predicted candidate genes.

For each of the eight network datasets (Figure 4-4) prediction of both the candidate genes affecting and those not affecting pathogenicity was made. The prediction was performed by mapping pathogenic 'seed' genes (Appendix B Table B- 2 to B-5) and 'seed' genes that have no effect on the pathogenicity (Appendix B Table B-6) respectively into the particular network. In addition, both prediction scenarios were applied in each network. The flowchart displayed in

Figure 4-4 illustrates the steps in prediction of both pathogenic genes and genes that do not affect pathogenicity. The results are summarised in Table 4-4 and Table 4-5.

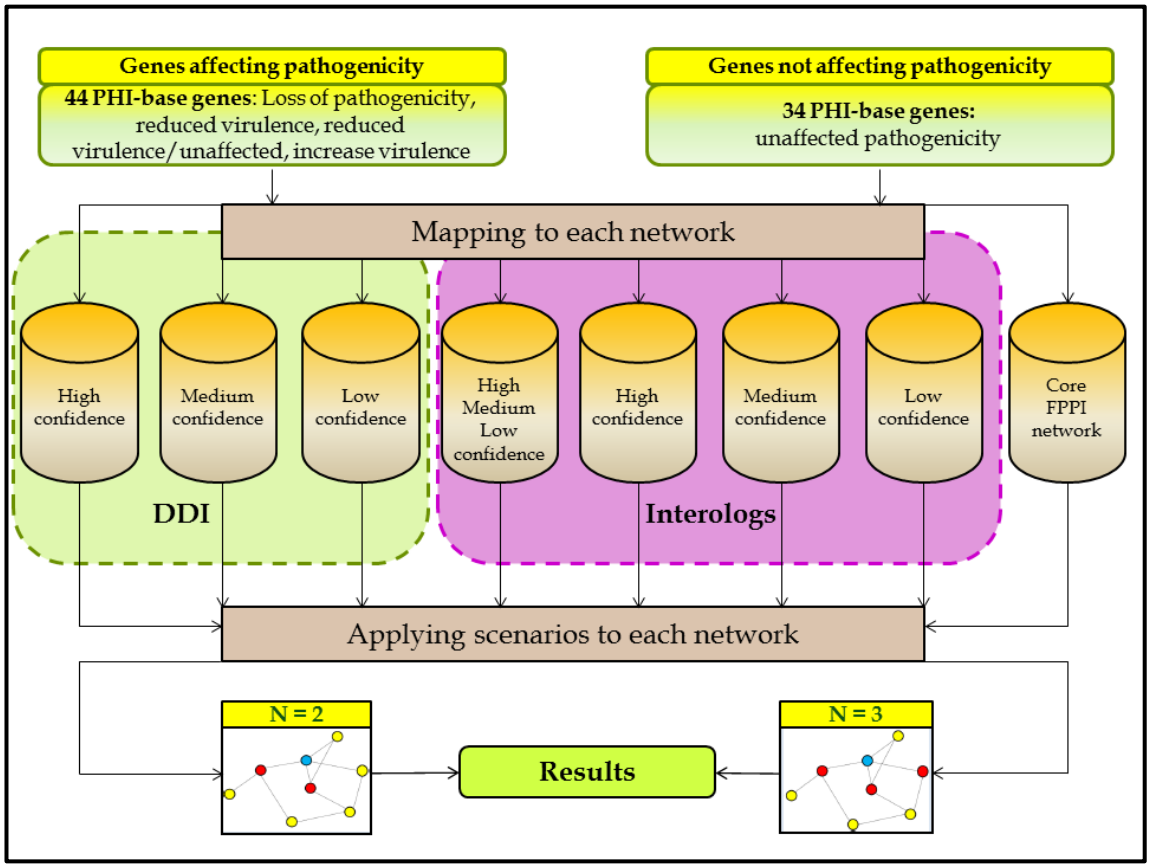


Figure 4-4 Flowchart for prediction of candidate genes in *F. graminearum*.
Core FPPI (*Fusarium graminearum* protein-protein interaction) network consists of high confidence DDI (domain-domain interaction) network and high confidence interologs network. Where N =2 and N=3 indicate minimum number of 'seed' genes (depicted in red) to be connected to the predicted candidate genes (depicted in blue).

Table 4-4 Summary of the 'seed' genes mapped to all datasets.

Network	No of Nodes in Network	'Seeds' mapped to network (C)	Pathogenic 'seeds' (A)	Non-pathogenic 'seeds' (B)
DDI_High Confidence	1716	16	11	5
DDI_Medium Confidence	2893	20	13	7
DDI_Low Confidence	4567	49	28	21
Interologs	3410	28	21	7
Interologs_High Confidence	1081	8	8	0
Interologs_Medium Confidence	2500	22	18	4
Interologs_Low Confidence	3207	28	21	7
DDI_Interologs_High Confidence (core network)	2610	24	19	5

Where (A): pathogenic 'seed' genes mapped to the network, (B): non-pathogenic 'seed' genes mapped to the network, where (A) + (B) = (C): the total number of 'seed' genes mapped to the given network.

Table 4-5 Summary of candidate genes prediction.

Network	No of Candidate Genes affecting pathogenicity		No of Candidate Genes not affecting pathogenicity	
	N = 2	**N = 3**	**N = 2**	**N = 3**
DDI_High Confidence	64	63	1	0
DDI_Medium Confidence	166	132	25	0
DDI_Low Confidence	629	387	83	57
Interologs	172	56	7	0
Interologs_High Confidence	1	0	0	0
Interologs_Medium Confidence	5	0	0	0
Interologs Low Confidence	142	40	6	0
DDI_Interologs_High Confidence (core network)	65	63	1	0

Shown are the number of candidate genes either affecting or not affecting pathogenicity. N defines the number of 'seed' genes that must be connected to the candidate gene (see Figure 4-3). DDI – domain-domain interaction network.

As seen from Table 4-5, the largest network prediction was made from the DDI datasets, i.e. based on pfam domains. This might be due to the fact that interolog datasets were built using *F. graminearum* orthologs from species that have not much in common with filamentous pathogenic fungi (see Table 2-2 in Chapter 2). Thus, most of the genes unique to pathogenic fungi and those that are specific to FG are missing in the interactome built via interologs approach.

4.3.2.1 Properties of networks created with different datasets

The main properties and structure measurements such as average clustering coefficient, average degree centrality for each network, and average centrality betweenness for the largest connected component of the analysed networks were calculated with the aid of NetworkX package and the results are presented in Table 4-6.

Investigating Table 4-6 further, differences can be observed in the average clustering coefficient and in the average degree centralities between the DDI and interologs networks. The significant differences in both network parameters between DDI high confidence and interologs high confidence networks were confirmed by Welch Two Sample t-test with p-value less than 0.05 at a 95% confidence interval.

Table 4-6 Main properties of the networks created with different datasets.

Dataset	Whole network			The largest connected component			
	Nodes	Edges	CCs	Nodes	Average clustering coefficient	Average Degree centrality	Average betweenness centrality
DDI_HighConfidence	1716	20019	244	128	0.8384	0.0136	0.0123
DDI_MediumConfidence	2893	43474	233	985	0.7086	0.0104	0.0053
DDI_LowConfidence	4567	105403	153	3516	0.7396	0.0101	0.0012
Interologs	3410	49079	23	3364	0.1929	0.0084	0.0006
Interologs_HighConfidence	1081	1726	102	817	0.2411	0.003	0.0075
Interologs_MediumConfidence	2500	11119	26	2444	0.0793	0.0036	0.0012
Interologs_LowConfidence	3207	36176	22	3163	0.1505	0.007	0.0006
DDI-InterologsHighConfidence	2610	21690	260	1293	0.6266	0.0064	0.0045

Where CCs- number of connected components.

Moreover, the two-sided Kolmogorov-Smirnov (KS) statistical test was used to compare the distribution of both network parameters between DDI high confidence and interologs high confidence networks. KS is a non-parametric statistical test that works on a cumulative distribution function. These two statistical tests revealed that there is a significant difference between the distribution of both clustering coefficient and degree centralities between DDI high confidence and interologs high confidence networks (p-value < 0.05 for both clustering coefficient and degree centralities comparison). Comparison of the average clustering coefficient between FPPI networks built with the DDI approach and those built with the interologs approach revealed that genes in an FPPI network built with the DDI approach are more likely to cluster together than genes in an FPPI network built with the interologs approach.

Furthermore, comparing the number of connected components within eight datasets listed in Table 4-6, it is interesting to note that the interologs dataset has the lowest number of connected components where the size of the largest connected component consists of the majority of the nodes presented in the interologs dataset. This might suggest that the spread of the information on the network that was built with interologs approach is more efficient, since starting from the node in the largest component it is possible to reach and share the information with the other nodes in the same component.

4.3.2.2 Prediction of *F. graminearum* pathogenic genes using the core FPPI network

The prediction of pathogenic genes in *F. graminearum,* based on the assumption that the unknown gene must be connected with at least two 'seed' genes to be treated as the candidate gene for pathogenicity, is illustrated in Figure 4-5. This prediction was made using FPPI high confidence DDI and FPPI high confidence interologs dataset (so-called 'core network'). In total 19 'seed' genes were mapped to this network, while the prediction was made only on eight 'seed' genes decorated in red (Figure 4-5). The prediction was calculated with aid of the NetworkX - python package and visualised using Ondex software (Köhler et al., 2006).

It is interesting to note that the separate small connected component where two 'seed' genes are connected to one predicted candidate (FGSG_06444). This predicted gene is a prime candidate for further biological experiments. This is because 'seed' gene: FGSG_03537 is a TRI5 gene responsible for mycotoxin production in *F. graminearum* and is specific to the *Fusarium* genus (Proctor et al., 1995). The second 'seed' gene: FGSG_10397 has no assigned protein function but it was proven to increase *F. graminearum* virulence when disrupted and tested in the wheat ear infection assay (Gardiner et al., 2009).

Furthermore, it was interesting to see if the predicted candidate genes are protein regulators or non-regulators, which in fact could give insight into the direction of the connection between 'seed' and predicted candidate genes. This information, however, was only available after the prediction was made.

Table 4-7 lists functional annotation of 'seed' genes used for prediction (depicted in red in Figure 4-5) of candidate genes for pathogenicity, whereas Table 4-8 itemises functional annotation of predicted candidate genes for pathogenicity (depicted in blue in Figure 4-5). The functional annotation of 'seed' and predicted candidate genes is based on the PHI-base version 4.2 (release date: 3rd October 2016). The information if the predicted candidate gene is a regulatory or a candidate for regulatory protein is taken from the study by Guo et al. (2016).

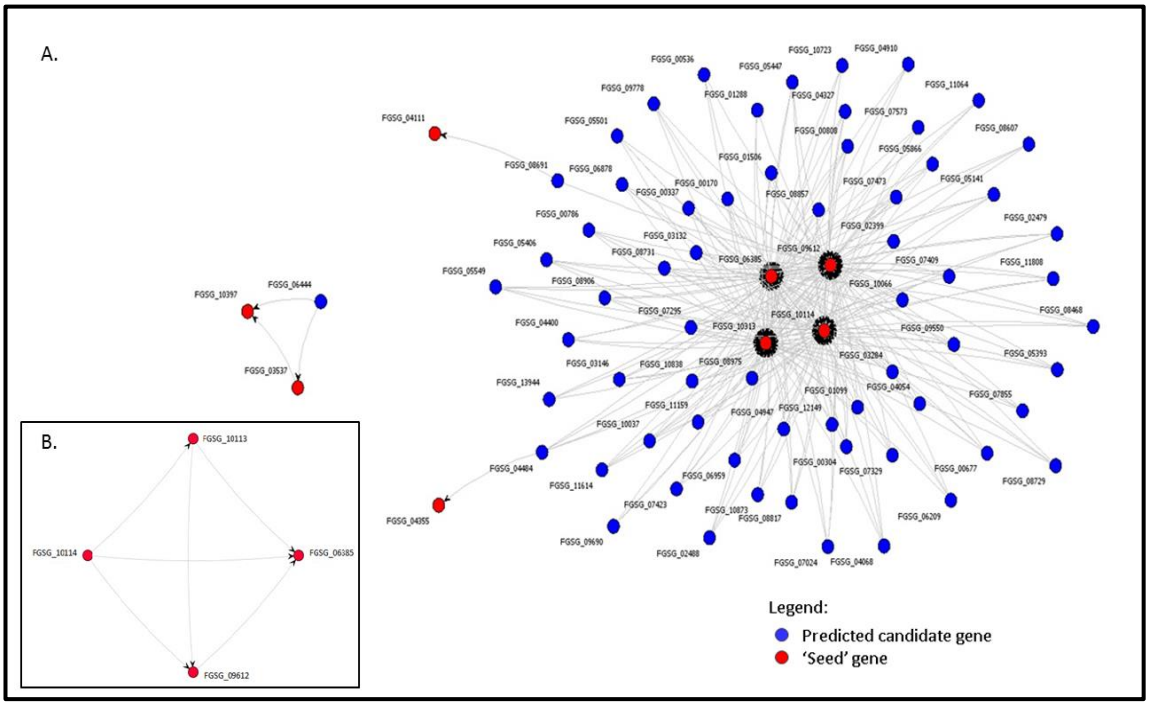


Figure 4-5 Prediction of candidate genes for pathogenicity in *F. graminearum* using the core FPPI network dataset.
A. Edges indicate the first direct neighbour of a predicted candidate gene in the FPPI network. Arrowheads are an artefact of a graph layout of the algorithm used for the drawing and directed towards the 'seed' gene participating in the prediction of a particular candidate gene (depicted in blue). The four 'seed' genes in the middle are FGSG_06385 (pathogenicity MAP kinase 1), FGSG_09612 (probable osmotic sensitive-2 protein (putative mitogen-activated protein - MAP kinase homolog)), FGSG_10313 (MAP kinase) and FGSG_10114 (probable RAS-2 protein).
B. Zoomed fragment of the drawing A. illustrates connections between four 'seed' genes: FGSG_10313, FGSG_06385, FGSG_09612 and FGSG_10114. Edges indicate the first direct neighbour, which is a 'seed' gene, of a 'seed' gene in the FPPI network, whereas arrowheads are an artefact of a graph layout of the algorithm used for the drawing and directed towards the 'seed' gene participating in the prediction of a particular candidate gene (here also a 'seed' gene).

Obviously, this information was not available during the time the prediction was made (April/May 2011) and it was added later to this chapter in 2016. When the additional gene annotation of predicted candidate genes became available, it was used to infer the direction of the interaction between the 'seed' gene and the predicted candidate gene. As all candidate genes are predicted to be protein regulators, either protein kinases or binding units. It is, however, difficult to speculate on which are the enzymes and which are the substrates. This makes it hard to infer a directional relationship between 'seed' genes and predicted candidate genes.

Table 4-7 Functional annotation of 'seed' genes in the core FPPI network

FGSG gene ID	PHI-base ID	Protein annotation*	Additional info**
FGSG_03537	PHI:44	trichodiene synthase	
FGSG_04111	PHI:2326	type 2C protein phosphatase	
FGSG_04355	PHI:2418	C-type cyclins	
	PHI:2419	C-type cyclins	
FGSG_06385	PHI:1189	protein kinase	candidate regulator
FGSG_09612	PHI:2327	mitogen-activated protein kinase	candidate regulator
FGSG_10114	PHI:861	Ras GTPase	regulator
FGSG_10313	PHI:1196	protein kinase	candidate regulator
FGSG_10397	PHI:2394	Conserved hypothetical protein	

*Annotation based on the information in PHI-base version 4.2 (release date: 3rd October 2016); **Information based on Guo's study (Guo et al., 2016).

Table 4-8 Functional annotation of predicted candidate genes in the core FPPI network.

FGSG gene ID	PHI-base ID	Protein annotation*	Additional info**
FGSG_00170	PHI:3834	probable rho3 protein	candidate regulator
FGSG_00304	PHI:3835	GTP binding	candidate regulator
	PHI:3836	Rho GTPases	candidate regulator
FGSG_00808	*PHI:4998*	GTP binding	candidate regulator
FGSG_01099		GTP-binding nuclear protein GSP1/Ran	candidate regulator
FGSG_01506	PHI:1251	protein kinase	candidate regulator
FGSG_02399	PHI:1239	protein kinase	candidate regulator
FGSG_02479		probable Ras-related protein Rab-6A	candidate regulator
FGSG_03146	PHI:1270	protein kinase	candidate regulator
FGSG_03284	PHI:1213	protein kinase	candidate regulator
FGSG_04068	PHI:3833	Rho GTPases	candidate regulator
FGSG_04327		Rab GTPases	candidate regulator
FGSG_04400		probable GTPase Rho	candidate regulator
FGSG_04947	PHI:1192	protein kinase	candidate regulator
FGSG_05141		GTP-binding protein ypt7	candidate regulator
FGSG_05406		protein kinase	candidate regulator
FGSG_05447		Rho GTPases	candidate regulator
FGSG_05501		probable VPS21 - GTP-binding protein	candidate regulator
FGSG_05549	PHI:1265	uncharacterised protein/ protein kinase activity	candidate regulator
FGSG_05866		hypothetical protein	
FGSG_06209		GTP-binding protein SAS1	regulator
FGSG_06444		conserved hypothetical protein	
FGSG_06959	PHI:1230	protein kinase	candidate regulator
FGSG_07024		Ras GTPases	candidate regulator
FGSG_07473		probable ras-related GTP-binding protein	candidate regulator
FGSG_07573		probable GTP-binding protein Drab11	candidate regulator
FGSG_08691	PHI:1190	protein kinase	candidate regulator
FGSG_08817		probable novel protein of ras superfamily KREV-1	regulator
FGSG_08857	PHI:3831	Rho GTPases	candidate regulator
FGSG_08975		septum-promoting GTP-binding protein 1	candidate regulator
FGSG_09550		related to GTP-binding protein Rab5c	candidate regulator
FGSG_10037	PHI:1201	protein kinase	candidate regulator
FGSG_10723		probable hymA gene	
FGSG_10838		related to transforming protein rho	candidate regulator
FGSG_10873		probable GTP-binding protein ypt1	candidate regulator
FGSG_11159		related to dis1-suppressing protein kinase dsk1	

Table 4-8 continues

FGSG gene ID	PHI-base ID	Protein annotation*	Additional info**
FGSG_11614	PHI:1278	protein kinase	candidate regulator
FGSG_11808		GTP-binding protein ypt5	regulator
FGSG_13944	PHI:1285	protein kinase	
FGSG_00337	PHI:1244	protein kinase	candidate regulator
FGSG_00536		conserved hypothetical protein	
FGSG_00677	PHI:1218	protein kinase	regulator
FGSG_00786	PHI:1249	protein kinase	regulator
FGSG_01288		casein kinase II subunit beta-2	candidate regulator
FGSG_02488	PHI:1269	protein kinase	candidate regulator
FGSG_03132	PHI:1266	protein kinase	candidate regulator
FGSG_04054	PHI:1221	protein kinase	candidate regulator
FGSG_04484	PHI:1175	protein kinase	candidate regulator
FGSG_04910		probable cyclin-dependent kinases regulatory subunit CKS1	candidate regulator
FGSG_05393	PHI:1223	protein kinase	candidate regulator
FGSG_06878	PHI:1183	protein kinase	candidate regulator
FGSG_07295	PHI:1194	protein kinase	candidate regulator
	PHI:3917	encoding the putative MAPKK	candidate regulator
FGSG_07329	PHI:1200	protein kinase	candidate regulator
FGSG_07409	PHI:1231	protein kinase	candidate regulator
FGSG_07423	PHI:1232	protein kinase	candidate regulator
FGSG_07855	PHI:1233	protein kinase	candidate regulator
FGSG_08468	PHI:1178	protein kinase	candidate regulator
FGSG_08607		probable casein kinase II beta subunit CKB1	candidate regulator
FGSG_08729	PHI:1240	protein kinase	candidate regulator
FGSG_08731	PHI:1235	protein kinase	candidate regulator
FGSG_08906	PHI:1215	protein kinase	candidate regulator
FGSG_09690		probable peptidylprolyl isomerase (FK506-binding protein homolog)	candidate regulator
FGSG_09778		probable transforming protein Ras-1	candidate regulator
FGSG_10066	PHI:1203	protein kinase	candidate regulator
FGSG_11064		related to glycine-rich RNA-binding protein	
FGSG_12149	PHI:1258	protein kinase	candidate regulator

Where FGSG genes highlighted in bold are *F. graminearum* genes predicted by Lysenko's study (Lysenko et al., 2013); *Annotation based on the information in PHI-base version 4.2 (release date: 3rd October 2016); ** Information based on Guo's study (Guo et al., 2016)**.**

Further analysis concentrates on one of the 'seed' gene's from the largest connected component, namely FGSG_10313. Comparison of this gene to the all connected candidate genes (first neighbours of the FGSG_10313 gene) with respect to MIPS functional categories (Appendix B Table B-7) and position of the predicted genes in the various cellular compartments (where specified) was performed. Information on subcellular localisation of a protein can help to predict the protein function. The information from WoLF PSORT analysis on the BROAD FG3 gene call,

provided by Dr John Antoniw from Rothamsted Research, was used to help in protein function prediction of candidate genes.

WoLF PSORT is a program for protein subcellular localisation prediction. It converts amino acids sequences into numerical vectors (numerical cell localisation), which are classified by simple k-nearest neighbour classifier. WoLF PSORT also organises proteins into more than 10 localisation sites including dual localisation for proteins moving between cytosol and nucleus (Horton et al., 2007).

Based on the information provided by Dr John Antoniw, the score equal/or greater than 17 was chosen in defining a cellular component for each *F. graminearum* gene connected to the 'seed' gene: FGSG_10313. The results are displayed in Table 4-9. Inspection of Table 4.9 highlights the lack of extracellular compartmentalised proteins within any of FG predicted pathogenicity genes.

Ondex (Köhler et al., 2006) software was used to display the first neighbours of the 'seed' gene FGSG_10313 *(MGV1*) with the information about the cellular component for predicted genes (where specified). In addition, the functional comparison between a 'seed' gene and predicted candidate genes was carried out using MIPS FunCat assignments. Thus, the predicted candidate genes that share six or more MIPS FunCat with the 'seed' gene are decorated with a triangle shaped node (Figure 4-6). *MGV1* (FGSG_10313) is a MAP kinase gene. MIPS FunCat of FGSG_10313 are metabolism (01), cell cycle and DNA processing (10), protein fate (14), cellular communications/signal transduction mechanism (30), cell rescue, defense and virulence (32), interactions with the cellular environment (34), cell fate (40), cell type differentiation (43), subcellular localization (70).

Table 4-9 Predicted candidate genes, first neighbours to FGSG_10313 'seed' gene, with MIPS FunCat and cellular compartment information.

No	Predicted gene	MIPS FunCat	Predicted cellular compartment	Highest WoLF PSORT score
1	FGSG_00170	**01**, **10**, 20, 16, **30**, **34**, **40**, 41, 42, **43**, **70**		
2	FGSG_00304	**01**, 16, **30**, **34**, **40**, 41, 42, **43**, **70**		
3	FGSG_00337	**01**, **14**, 16, 18, **30**, **34**, **40**, **70**		
4	FGSG_00536	11, 16, **70**		
5	FGSG_00677	**01**, **10**, 11, **14**, 16, 18, **30**, **32**, **34**, **40**, 41, 42, **43**, **70**	nucl	**17**
6	FGSG_00786	**01**, **14**, 16, 18, **30, 70**		
7	FGSG_00808	16, 20, **30**,	nucl	**15**
8	FGSG_01099	**01**, **10**, 11, **14**, 16, 20, **30**, 42, **70**	cyto	**17**
9	FGSG_01288			
10	FGSG_01506	**01**, **10**, **14**, 16, 18, **30**, **32**, 41, **70**		
11	FGSG_02399		nucl	**21.5**
12	FGSG_02479	16, 20, **30**, **70**, 98	nucl	**12.5**
13	FGSG_02488	**30**	nucl	**13.5**
14	FGSG_03132		cyto, nucl	**9.5**
15	FGSG_03146		nucl	**10.5**
16	FGSG_03284		cyto	**11.5**
17	FGSG_04054	**01**, **10**, **14**	nucl	**14**
18	FGSG_04068	**01**, 16, **30**, **34**, **40**, 41, 42, **43**, **70**	cyto	**17.5**
19	FGSG_04327	**01**, 16, 20, **30**, **70**	cyto	**12**
20	FGSG_04400	**01**, **10**, 16, **30**, **34**, **40**, 41, 42, **43**, **70**	cyto, mito	**11.8**
21	FGSG_04484	**01**, **10**, 11, **14**, 16, 18, **34**, **43**, **70**	mito	**17.5**
22	FGSG_04910	**10**	nucl	**17**
23	FGSG_04947	**01**, **10**, **14**, 16, **30**, **34**, **40**	cysk	**15**
24	FGSG_05141	16, 20, **30**, **70**	mito	**12.5**
25	FGSG_05393	**01**, 02, **10**, 11, **14**, 16, 18, **30**, **34**, **40**, 41, **70**, 77	nucl	**11**
26	FGSG_05406	**01**, **10**, **14**, 16, 18, **30**, **70**	nucl	**11.5**
27	FGSG_05447	**01**, **10**, 16, 18, 20, **30**, **32**, **34**, **40**, 41, 42, **43**, **70**	cyto	**21.5**
28	FGSG_05501	**01**, **14**, 16, 20, **30**, **70**	cyto	**15.5**
29	FGSG_05549		nucl	12.5
30	FGSG_05866		cyto	16
31	FGSG_06209	20, **30**	cyto	8.5
32	FGSG_06878	**14**, 16, **30**, **34**, **40**	cyto	15.5

Table 4-9 continues

No	Predicted gene	MIPS FunCat	Cellular compartment	Highest WoLF PSORT score
33	FGSG_06959	**01**, **10**, **14**, 16, 18, **30**, 42, **70**	nucl	19.5
34	FGSG_07024	**14**, 16, 18, **30**, **40**, 42, **43**, **70**, 73	mito	10
35	FGSG_07295		cyto	13.5
36	FGSG_07329	**01**, **14**, 16, 18, **30**, **40**, 45, **70**	nucl	17
37	FGSG_07409		nucl	17
38	FGSG_07423	**01**, **10**, 11, **14**, 16, 18, **30**, **70**	mito	13
39	FGSG_07473	**01**, 11, 16, 20, **30**, **70**	mito	18.5
40	FGSG_07573	**01**, 16, 20, **30**, **70**	cyto, nucl	13.8
41	FGSG_07855	**01**, **10**, 11, **14**, **30**, **32**, **34**, **40**, 42, **43**	nucl	16.6
42	FGSG_08468	**01**, 02, **10**, 11, **14**, 16, 18, **30**, **34**, **40**, **43**, **70**	cyto	11.5
43	FGSG_08607	**01**, **10**, 11, **14**, 16, **30**, **32**, **34**, **40**, 42, **43**	nucl	12
44	FGSG_08729	**01**, 11, **14**, 16, **30**, **43**, **70**	cyto	13.5
45	FGSG_08731		nucl	11
46	FGSG_08817	**01**, **10**, 16, **30**, **34**, 41, **43**	nucl	14.5
47	FGSG_08857	**01**, **10**, 16, 18, 20, **30**, **32**, **34**, 36, **40**, 41, 42, **43**, **70**	nucl	12.5
48	FGSG_08906		nucl	11.5
49	FGSG_08975		nucl	11
50	FGSG_09550	**01**, 16, 20, **30**, 42, **70**	mito	15
51	FGSG_09690	**10**, 12, **14**, 16, **30**, **32**, **34**, 36, **70**, 75	mito	23
52	FGSG_09778	**01**, **10**, 16, **30**	cyto	10
53	FGSG_10037	**01**, **14**, 16, **30**, **43**, **70**, 98	cyto	8
54	FGSG_10066	**14**, **30**, **40**, **70**	nucl	13
55	FGSG_10723	**10**, 11, 16, **40**, **43**	mito	13
56	FGSG_10838	**30**	nucl	13.5
57	FGSG_10873	**10**, 11, **14**, 16, 20, **30**, **32**, 42, **70**	cyto, nucl	11.83
58	FGSG_11064		nucl	9.5
59	FGSG_11159		cyto	21
60	FGSG_11614		mito	10.5
61	FGSG_11808		mito	12.5
62	FGSG_12149		cysk	10
63	FGSG_13944		cyto	11

Highlighted in bold are MIPS FunCat in common with the FGSG_10313 'seed' gene. These are Metabolism (01), Cell cycle and DNA processing (10), Protein fate (14), Cellular communications/signal transduction mechanism (30), Cell rescue, defense and virulence (32), Interactions with the cellular environment (34), Cell fate (40), Cell type differentiation (43), Subcellular localization (70). Cellular compartments are nucl – nucleus, cysk – cytoskeleton, cyto – cytosol, mito – mitochondria**.**

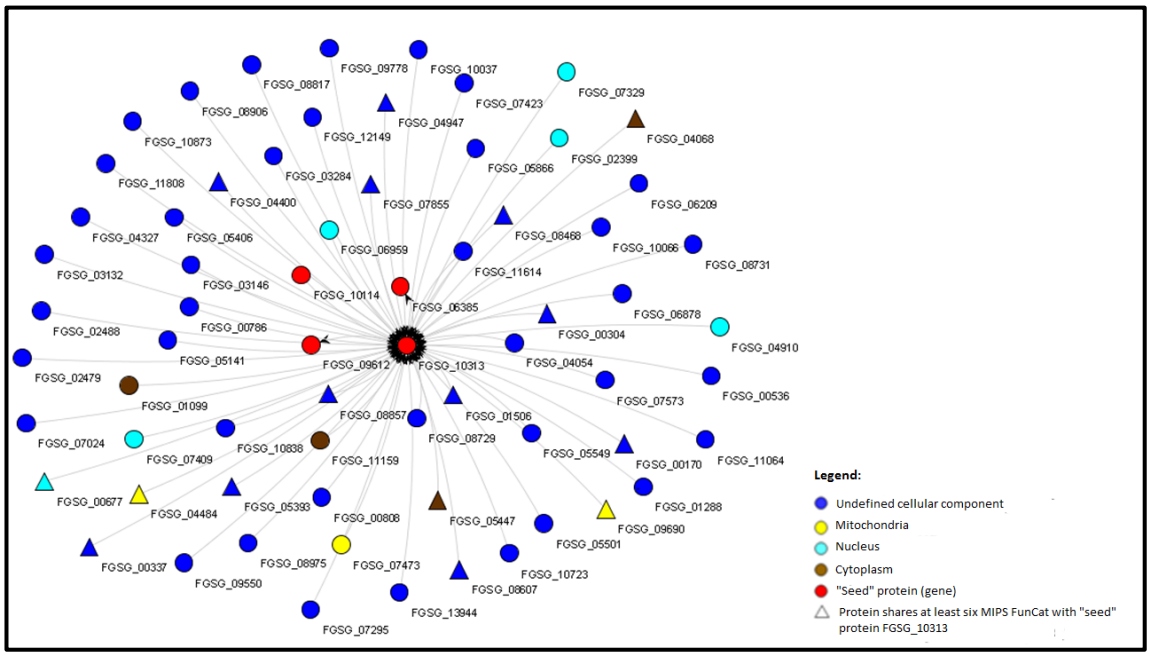


Figure 4-6 Characteristic features of predicted candidate genes connected to MGV1 gene (FGSG_10313) in the core FPPI network.
Edges indicate a connection to the first direct neighbours of *MGV1* gene in the FPPI network, while arrowheads direct towards the 'seed' gene participating in the prediction of a candidate gene for pathogenicity in *F. graminearum*. The candidate genes are depicted in the colour associated with cellular compartment calculated with the aid of the WoLF PSORT algorithm, when defined or depicted in blue colour otherwise. Also, the candidate genes that share six or more MIPS FunCat functions with the 'seed' gene FGSG_10313 are illustrated as triangle nodes.

With reference to Table 4-9 it is interesting to note that each of the predicted candidate genes shares at least one MIPS FunCat protein function (where assigned) with the 'seed' gene FGSG_10313. Moreover, inspecting further Table 4-9 and Figure 4-6, a quarter of the predicted candidate genes share six or more MIPS FunCat protein functions with the 'seed' gene FGSG_10313. This result suggests that the 'seed' gene used for the prediction and the predicted candidate genes share similar protein functions. Thus, if the 'seed' gene, namely FGSG_10313 has been experimentally proven to affect pathogenicity, the candidate genes directly connected to this 'seed' gene in this study and sharing six or more MIPS FunCat protein functions with the 'seed' gene are quite likely to influence pathogenicity. In other words, this finding supports the prediction of candidate genes for the pathogenicity in this study.

4.3.2.3 Properties of the core network

The core network is defined as the network consisting of interactions from high confidence interologs and high confidence DDI networks. In this section the core network properties such as

degree centralities and clustering coefficient were separately calculated for four different gene groups. The ‘seed’ genes were divided into the ‘predictive seed’ gene set (group 1) and ‘non-predictive other seed’ gene set (group 2). The ‘predictive seed’ genes are those that contribute to the prediction of the pathogenic genes in the core network (a scenario depicted in Figure 4-3 A). The ‘non-predictive other seed’ genes are mapped to the core network but do not contribute to the prediction of candidate pathogenic genes. The remaining two gene groups are the predicted pathogenic and non-predicted gene sets.

The splitting of the ‘seed’ genes into two groups made it possible to test if any similarities in the network parameters between the gene groups exist. Do nodes of predicted candidate genes have similar network properties in the network compared to the nodes of ‘seed’ genes that took part in their prediction? Can those parameters be used to validate the prediction of the candidate genes for pathogenicity? The outcome of this analysis is displayed in Table 4-10 showing the average values of the main network properties of all four gene groups.

Table 4-10 The average values of the main properties of the core network with respect to a different group of nodes in the network.

Genes	Degree	Degree centralities	Clustering coefficient
‘non-predictive other seed’ genes	13	0.0050	0.4706
‘predictive seed’ genes	35	0.0132	0.7241
Predicted pathogenic genes	62	0.0237	0.8440
Non-predicted genes	15	0.0059	0.6214

Where: ‘non-predictive other seed’ genes are the ‘seed’ genes mapped to the network but did not take part in the prediction of the candidate pathogenic genes, ‘predictive seed’ genes are the ‘seed’ genes based on which the prediction of candidate pathogenic genes was made (as per scenario depicted in Figure 4-3 A).

From Table 4-10 it can be observed that clustering coefficient and degree centralities of predicted pathogenic genes are closer in value to those of ‘predictive seed’ genes used in the prediction of pathogenic genes.

Degree-centrality distribution comparison

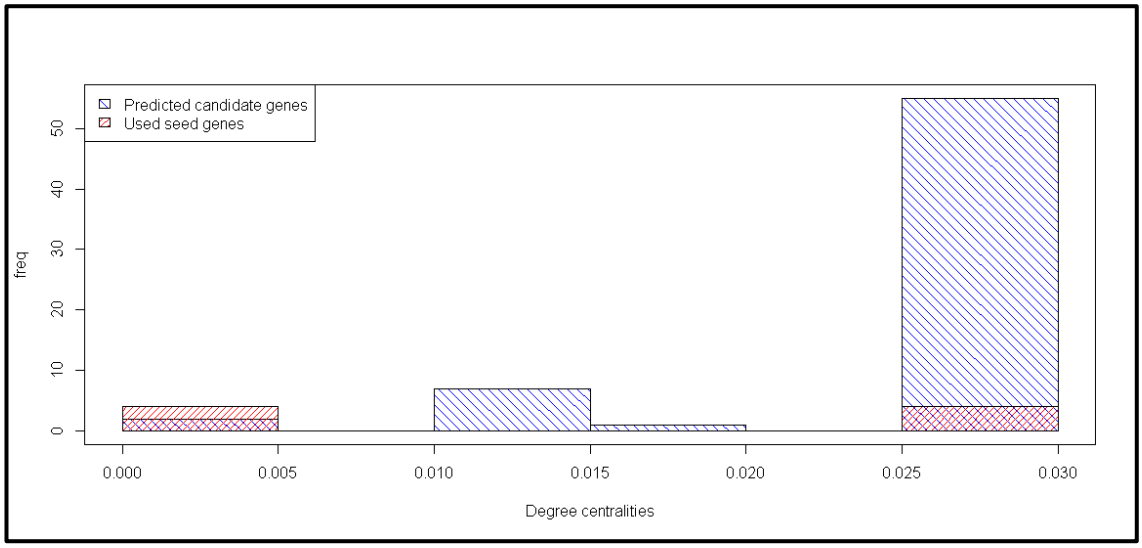


Figure 4-7 Comparison of degree-centrality distribution.
The comparison was made between degree-centrality distribution of predicted candidate genes and degree-centrality distribution of 'predictive seed' genes used for the prediction.

Figure 4-7 illustrates the degree-centrality distribution comparison between 'predictive seed' genes and predicted candidate genes sets. Both sets are not normally distributed and non-parametric statistical tests were performed to compare these distributions.

Firstly, Wilcoxon-Mann-Whitney rank sum test (equivalent to Wilcoxon Rank Sum test with continuity correction in R package) was used to compare averages of these two independent genes sets (Table 4-11 A). As a result of the test on 5% confidence level, no significant difference in mean values was observed for 'predictive seed' genes and predicted pathogenic genes (p-value = 0.09612, Table 4-11 A.). Furthermore, no significant difference (p-value > 0.05) in mean values was also noticed while testing two other pairs of gene sets, namely 'predictive seed' genes and non-predicted genes (p-value = 0.3532, Table 4-11 A.), as well as 'non-predictive other seed' genes and non-predicted gene pair sets (p-value = 0.08032, Table 4-11 A.). The lack of significant differences between average degree centrality of 'predictive seed' genes and average degree centrality of non-predicted genes can be explained by the large disproportion in sample sizes for compared gene sets.

Finally, a significant difference between mean values was detected while comparing 'non-predictive other seed' genes set with predicted pathogenic genes set. The initial findings from the

Wilcoxon-Mann-Whitney rank sum test were confirmed by the stronger non-parametric statistical test for comparing two independent distributions, namely the Kolmogorov-Smirnov (KS) test (Table 4-11 B).

Table 4-11 Degree-centrality distribution comparison.

A. Wilcoxon-Mann-Whitney rank sum test (an equivalent to Wilcoxon Rank Sum Test with continuity correction in R package)

Dataset		Sample size		Median		W	p-value
1	2	1	2	1	2		
'non-predictive other seed' genes	Predicted pathogenic genes	11	65	0.0008	0.0253	74.5	0.000019
'non-predictive other seed' genes	Non-predicted genes	11	2526	0.0008	0.0019	9681	0.080320
'predictive seed' genes	Predicted pathogenic genes	8	65	0.0132	0.0253	168.5	0.096120
'predictive seed' genes	Non-predicted genes	8	2526	0.0132	0.0019	12010	0.353200

B. Kolmogorov-Smirnov test

Dataset		Sample size		Median		D	p-value
1	2	1	2	1	2		
'non-predictive other seed' genes	Predicted pathogenic genes	11	65	0.0008	0.0253	0.8322	0.000004
'non-predictive other seed' genes	Non-predicted genes	11	2526	0.0008	0.0019	0.3076	0.2511
'predictive seed' genes	Predicted pathogenic genes	8	65	0.0132	0.0253	0.4692	0.08684
'predictive seed' genes	Non-predicted genes	8	2526	0.0132	0.0019	0.4272	0.1089

Where: 'non-predictive other seed' genes are the 'seed' genes mapped to the network but did not take part in the prediction of the candidate pathogenic genes, 'predictive seed' genes are the 'seed' genes based on which the prediction of the candidate pathogenic genes was made (as per scenario depicted in Figure 4-3 A).

Clustering coefficient distribution comparison

Figure 4-8 illustrates the clustering coefficient distribution comparison between 'predictive seed' genes and predicted candidate genes sets. As both sets are not normally distributed, non-parametric statistical tests were performed to compare these distributions. Firstly, Wilcoxon-Mann-Whitney rank sum test (equivalent to Wilcoxon Rank Sum test with continuity correction in R package) was used to compare averages of these two independent genes sets (Table 4-12 A.).

As a result of the test on 5% confidence level, no significant difference in mean values was observed for 'predictive seed' genes and predicted pathogenic genes (p-value = 0.9711).

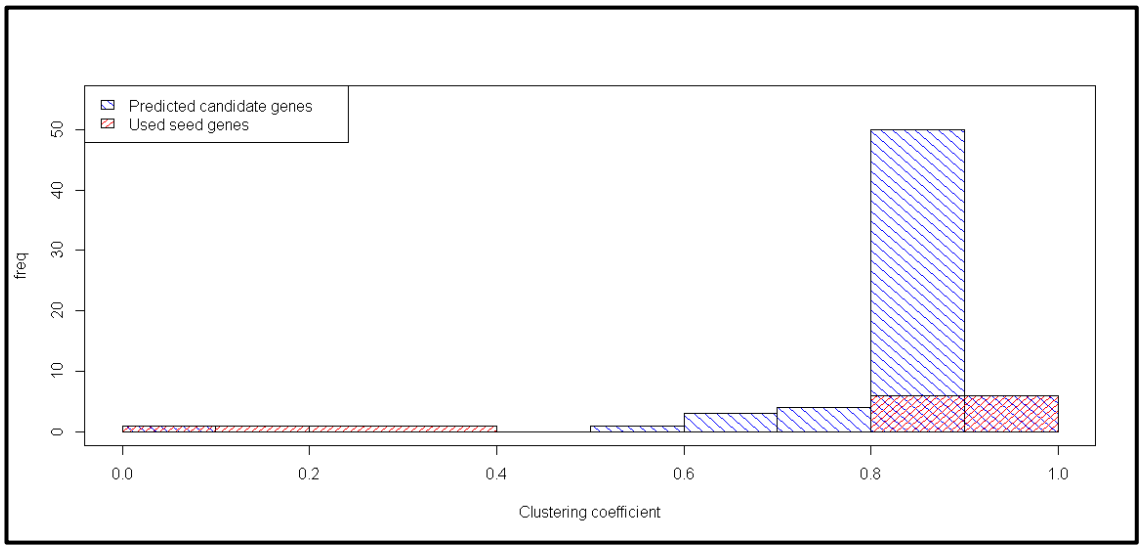


Figure 4-8 Comparison of clustering coefficient distribution.
The comparison was made between clustering coefficient distribution of predicted candidate genes and clustering coefficient distribution of 'predictive seed' genes used for the prediction.

Moreover, no significant difference in average clustering coefficient was detected for the remaining three pairs of tested gene sets, namely 'non-predictive other seed' genes and predicted pathogenic genes pair (p-value = 0.2088), 'non-predictive other seed' genes and non-predicted genes pair (p-value = 0.2608), as well as 'predictive seed' genes and non-predicted genes pair (p-value = 0.8038).

Analogically to degree centrality parameter, the strongest similarity in mean values between 'predictive seed' genes and non-predicted genes can be explained by the very small sample size of 'predictive seed' genes (8) comparing to a very large sample size of the non-predicted genes set (2526). The initial finding from the Wilcoxon-Mann-Whitney rank sum test was confirmed by the stronger non-parametric statistical test for comparing two independent distributions, namely KS test (Table 4-12 B.). However, in the KS test, a weak significant difference was observed between distributions of 'non-predictive other seed' genes and predicted pathogenic genes (p-value = 0.0101).

Table 4-12 Clustering coefficient distribution comparison.

A. Wilcoxon-Mann-Whitney rank sum test (an equivalent to Wilcoxon Rank Sum Test with continuity correction in R package)

Dataset		Sample size		Median		W	p-value
1	2	1	2	1	2		
'non-predictive other seed' genes	Predicted pathogenic genes	11	65	0.3333	0.8716	273.5	0.20880
'non-predictive other seed' genes	Non-predicted genes	11	2526	0.3333	1.0000	11386	0.26080
'predictive seed' genes	Predicted pathogenic genes	8	65	0.8587	0.8716	257.5	0.97110
'predictive seed' genes	Non-predicted genes	8	2526	0.8587	1.0000	9631.5	0.80380

B. Kolmogorov-Smirnov test

Dataset		Sample size		Median		D	p-value
1	2	1	2	1	2		
'non-predictive other seed' genes	Predicted pathogenic genes	11	65	0.3333	0.8716	0.53007	0.01012
'non-predictive other seed' genes	Non-predicted genes	11	2526	0.3333	1.0000	0.19866	0.78040
'predictive seed' genes	Predicted pathogenic genes	8	65	0.8587	0.8716	0.23462	0.82780
'predictive seed' genes	Non-predicted genes	8	2526	0.8587	1.0000	0.28761	0.52440

Where: 'non-predictive other seed' genes are the 'seed' genes mapped to the network but did not take part in the prediction of the candidate genes, 'predictive seed' genes are the 'seed' genes based on which the prediction of candidate pathogenic genes was made (as per scenario depicted in Figure 4-3 A).

4.3.3 Exploring the genomic location of the genes predicted in core FPPI network

The previous study by Cuomo *et al.* (2007) on the assembled genome of *F. graminearum* aligned to the available genetic map identified distinct regions within chromosomes. These sub-regions possess the highest single nucleotide polymorphism (SNP) density and were also associated with the highest genetic recombination rate (deep red colour on the chromosomes, Figure 4-9). Other regions of the genome had very low / no SNPs and very low / no genetic recombination (blue to white colour on the chromosomes, Figure 4-9). Moreover, genes solely expressed during plant infection were found to reside preferentially within the high diversity regions of the genome (Cuomo et al., 2007).

Due to the high recombination events within these regions it is highly possible that genes present within these regions are more specific to *F. graminearum*, while genes present in lower recombination regions (light blue, Figure 4-9) are common among the *Fusarium* genus and other

plant pathogenic fungi. Thus, a further step in the analysis was to map the predicted candidate genes (listed in Table 4-8) together with 19 pathogenic 'seed' genes (from FPPI core network) to the *F. graminearum* genome. This was performed with the aid of OmniMapFree (Antoniw et al., 2011) and the results are displayed in Figure 4-9.

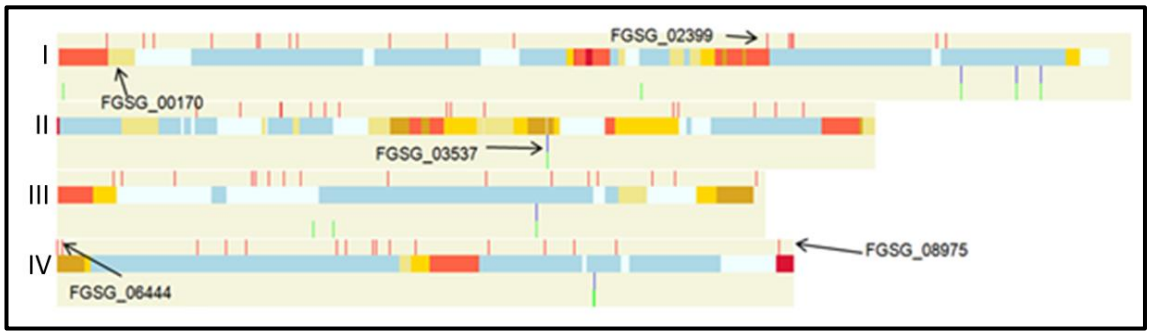


Figure 4-9 *F. graminearum* genome with mapped predicted candidate genes and 'seed' genes. The diagram displays the four FG chromosomes indicated by Roman numerals. The predicted pathogenic genes are depicted as red vertical bars in track 1. Recombination frequency in a cross between the sequenced strain PH-1 and a second USA strain called MN00-676 is shown in track 2 as a colour code, where recombination frequency increases through the colours: (low) azure, light blue, khaki, gold, golden red, tomato and crimson (high). The absolute recombination frequency ranges from zero to >8 centimorgan (cM) between consecutive genetic markers. For each chromosome, blue coloured vertical bars in the track 3 locate the FG genes coding for known pathogenic genes ('predictive seed' genes) mapped to the network, based on which prediction of pathogenic genes were made Row 4 green vertical lines represent known pathogenic genes mapped to the network ('non-predictive other seed' genes, not used for the prediction of candidate genes).

Further inspection of the Figure 4-9 revealed that most of the predicted genes reside within lower recombination regions. However, there are three genes, namely FGSG_00170, FGSG_02399, and FGSG_08975 within the highest recombination regions of the genome. FGSG_02399 was predicted to be a serine/threonine-protein kinase, FGSG_08975 was predicted to be a GTP binding protein and both are known intracellular signalling pathway components.

Additionally, gene FGSG_06444 connected to TRI5 gene (FGSG_03537) responsible for toxin production in *F. graminearum* (Figure 4-5) is positioned in the high region of recombination towards the end of chromosome IV.

4.3.4 Discussion

Inspired by the previous study (Liu et al., 2010) and using previously predicted protein-protein interactome for *F. graminearum* (Zhao et al., 2009), potential candidate genes responsible for pathogenicity in this economically important plant pathogenic fungus were predicted. To compare with the previous study (Liu et al., 2010), combining high confidence DDI and high confidence interologs network was used as the interactome onto which we mapped known for pathogenicity *F. graminarum* genes present in PHI-base version 3.2.

Although assuming that the same interactome dataset as per Liu *et al.* (2010) study was used, the number of proteins (nodes) and interactions (edges) within our network differs considerably. Furthermore, in contrast to the previous study careful dissection of all *F. graminearum* genes, present in PHI-base version 3.2, were performed into two groups: those that were proven to influence the pathogenicity and those that were proven not to influence the pathogenic process. Then, when predicting candidate genes for the pathogenicity only the genes that influenced the pathogenicity process were mapped into the network.

The differences in the prediction of pathogenic genes between Liu *et al.* (2010) study and our study are listed in Table 4-13. As we can observe from Table 4-13, 'predictive seed' genes on the base of which prediction was made in Liu *et al.* (2010) study are considerably different to our 'predictive seed' genes mapped to the network. It seems that there are only three 'seed' genes in common between both predictions. In our study, the 'seed' gene FGSG_09903 (*STE7*) is replaced by FGSG_10114 (*RAS2*) in the larger connected component (see Figure 4-5) and compare to Liu *et al.* (2010) prediction (Appendix B Figure B-1). However, inspection of the detailed intracellular signalling pathway map given in Figure 4-10 reveals that both 'seed' genes are part of the same MAPK signalling cascade.

Further investigation of the gene FGSG_09903 revealed that this gene should not be mapped to the combined DDI and interologs high confidence network (the core FPPI network) as it is only interacting with one other gene, which is a 'seed' gene FGSG_06385, within the interologs fraction of the core network. That is why this gene was not mapped to our pathogenic network.

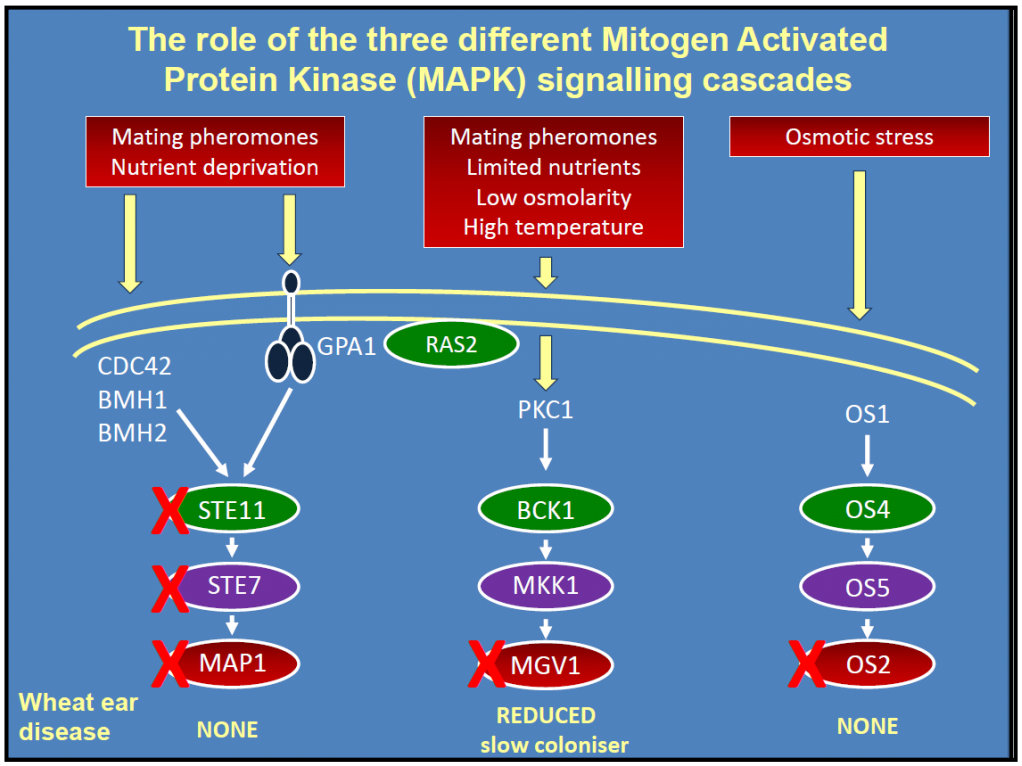


Figure 4-10 MAPK signalling cascades activated in *F. graminearum* during wheat ear infection.
Red crosses indicate mutation/silencing of genes.

Additionally, three 'predictive seed' genes: FGSG_10313 (*MGV1*), FGSG_09612 (*HOG1/Os-2*) and FGSG_06385 (*MAP1*), which were mapped in both our and Liu *et al.* (2010) pathogenic networks, are crucial genes that take part in MAPK signalling cascade.

As seen from Figure 4-10 (kindly provided by Dr Martin Urban, Rothamsted Research), the independent elimination of those genes from the signalling cascade leads to a loss in pathogenicity or reduced virulence in *F. graminearum* in all three cases. Thus, there is a high probability that predicted candidate genes, which are connected to these 'predictive seed' genes, are part of the same three MAPK signalling cascades and / or operate immediately upstream or downstream of these three important phosphorylation relays.

Further investigation of candidate genes for pathogenicity predicted by the previous study (Liu et al., 2010) reveals that gene FGSG_00838 (highlighted in yellow in Table 4-13) also should not be considered as a potential candidate gene for pathogenicity as it is only connected to the following

Table 4-13 Comparison of our prediction to prediction made by Liu's study (Liu et al., 2010).

No	Predicted Candidate genes for pathogenicity			Seed genes, mapped to the network, on basis of which prediction was made		
	In common	Liu's study only	This study only	In common	Liu's study only	This study only
1	**FGSG_00337**	FGSG_00760	FGSG_00170	FGSG_10313	FGSG_09903	FGSG_10397
2	**FGSG_00536**	FGSG_00838	FGSG_00304	FGSG_09612	FGSG_05484	FGSG_03537
3	**FGSG_00677**	FGSG_01338	FGSG_00808	FGSG_06385	FGSG_09197	FGSG_04355
4	**FGSG_00786**	FGSG_02584	FGSG_01099		FGSG_04104	FGSG_04111
5	FGSG_01288	FGSG_02648	**FGSG_01506**		FGSG_05535	FGSG_10114
6	**FGSG_02488**	FGSG_02795	FGSG_02399		FGSG_09614	
7	FGSG_03132	FGSG_04286	FGSG_02479		FGSG_09280	
8	**FGSG_04054**	FGSG_05038	**FGSG_03146**			
9	**FGSG_04484**	FGSG_05698	FGSG_03284			
10	**FGSG_04910**	FGSG_05737	FGSG_04068			
11	**FGSG_05393**	FGSG_06266	FGSG_04327			
12	**FGSG_06878**	FGSG_07335	**FGSG_04400**			
13	FGSG_07295	FGSG_09271	**FGSG_04947**			
14	**FGSG_07329**	FGSG_09660	FGSG_05141			
15	FGSG_07409	FGSG_09870	**FGSG_05406**			
16	**FGSG_07423**	FGSG_09988	FGSG_05447			
17	**FGSG_07855**	FGSG_10251	FGSG_05501			
18	**FGSG_08468**	FGSG_10804	FGSG_05549			
19	**FGSG_08607**	FGSG_10822	FGSG_05866			
20	**FGSG_08729**	FGSG_10894	**FGSG_06209**			
21	FGSG_08731	FGSG_11878	FGSG_06444			
22	FGSG_08906	FGSG_11979	**FGSG_06959**			
23	**FGSG_09690**		FGSG_07024			
24	**FGSG_09778**		FGSG_07473			
25	**FGSG_10066**		FGSG_07573			
26	**FGSG_11064**		**FGSG_08691**			
27	FGSG_12149		FGSG_08817			
28			**FGSG_08857**			
29			FGSG_08975			
30			FGSG_09550			
31			FGSG_10037			
32			FGSG_10723			
33			FGSG_10838			
34			**FGSG_10873**			
35			FGSG_11159			
36			**FGSG_11614**			
37			FGSG_11808			
38			**FGSG_13944**			

Where genes highlighted in bold were also predicted by Lysenko et al. (2013); gene highlighted in yellow - incorrectly predicted candidate gene for pathogenicity in Liu's study (Liu et al., 2010), as only connected to one pathogenic 'seed' gene; genes highlighted in grey - genes that are less likely to be candidates for pathogenicity as they are also connected to the 'seed' gene that was proven to have no effect on the pathogenicity (PHI-base version 3.2, release date 14th Dec 2009).

'seed' genes: FGSG_09614 and FGSG_09280 (see Appendix B Figure B-1) where FGSG_09280 gene does not exist in either version 3.1 or version 3.2 of PHI-base.

Additionally, 'seed' gene FGSG_05535, mapped to the pathogenic network by Liu *et al.* (2010) should not be taken into account as it was shown not to have an effect on the pathogenicity (see Appendix B Table B-6). Thus, predicted candidate genes highlighted in grey in Table 4-13 are less likely to be candidates for pathogenicity because these are connected to 'seed' genes that have been proven to affect pathogenicity (FGSG_04104 and FGSG_09614), as well as to the 'seed' gene that showed no effect on the pathogenicity process (FGSG_05535) (see Appendix B Table B-5 and Table B-6). Therefore, when the predicted pathogenic network by Liu *et al.* (2010) is redrawn to only contain the experimentally proven *F. graminearum* pathogenicity and virulence genes, the number of predicted pathogenic genes shrinks to 41.

4.4 Prediction of candidate genes in *Magnaporthe oryzae* – a pilot study

4.4.1 Using network for prediction

In this study the predicted *Magnaporthe oryzae* (MO) interactome dataset was downloaded from the supplementary materials of He's study (He et al., 2008). In order to find candidate genes two scenarios (Figure 4-3 A. and B.) were applied in each examination. The first scenario (Figure 4-3 A) applied by Liu et al. (2010), assumes that the unknown gene is the candidate gene if it is connected to at least two 'seed' genes, while in the second scenario (Figure 4-3 B) the unknown gene is the candidate gene if it is connected to three 'seed' genes. A 'seed' gene is defined as a MO gene that has been experimentally tested and shown to have a role in pathogenic properties or was shown not to have a role in pathogenicity. All 'seed' genes were dissected according to their phenotype (see Appendix B Tables B-8 to B-11). The flowchart displayed in Figure 4-11 demonstrates the steps in prediction of both pathogenic genes and genes that do not affect pathogenicity. The results are summarised in Table 4-14.

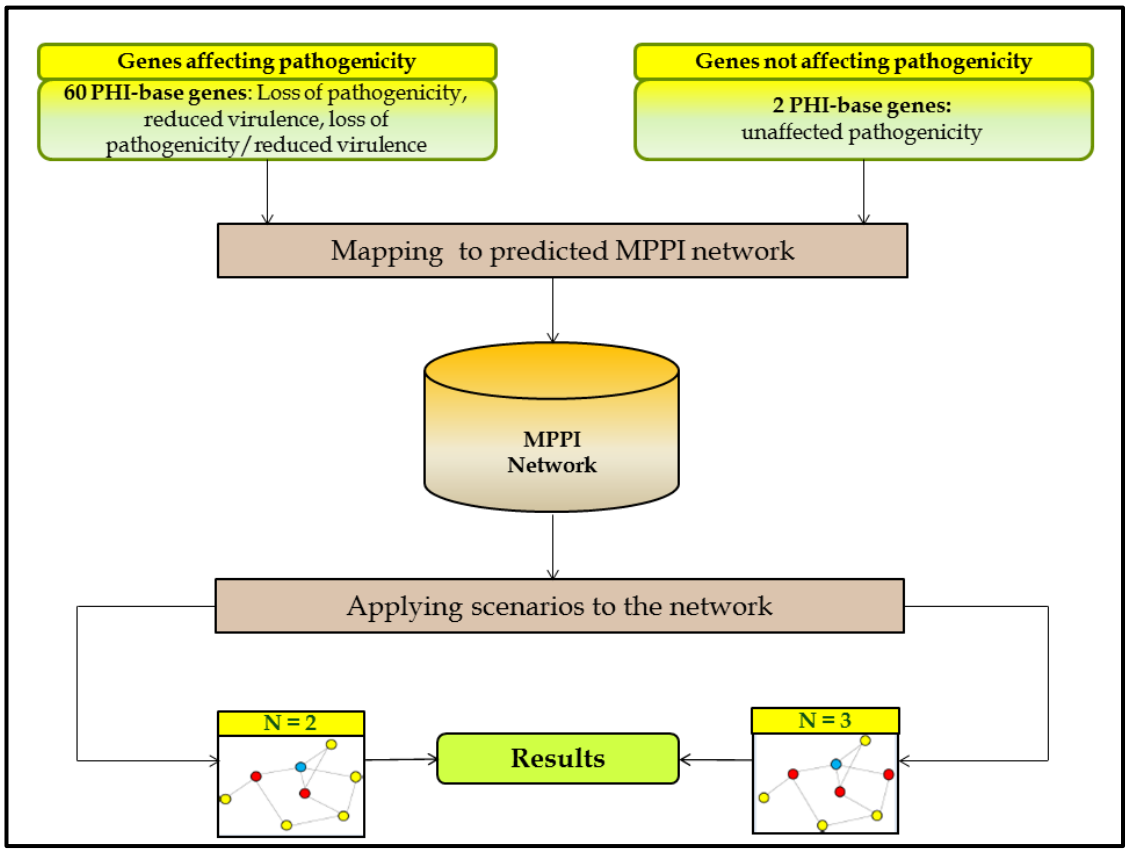


Figure 4-11 Flowchart for prediction of candidate genes in *M. oryzae*.
Where MPPI – *Magnaporthe oryzae* protein-protein interaction network, N = 2 and N=3 indicate a minimum number of 'seed' genes (depicted in red) to be connected to the predicted candidate (depicted in blue).

Table 4-14 Summary of prediction of candidate genes in *M. oryzae* in MPPI network.

MPPI network		
Number of nodes	3016	
'Seeds' mapped to the network (C)	24	
Pathogenic 'seeds' (A)	23	
Non-pathogenic 'seeds' (B)	1	
Number of candidate genes affecting pathogenicity	N = 2	8
	N = 3	1
Number of candidate genes not affecting pathogenicity	N = 2	0
	N = 3	0

Shown is the number of candidate genes either affecting or non-affecting pathogenicity. N defines the number of 'seed' genes that must be connected to the candidate gene (see Figure 4-3). Where: (A): pathogenic 'seed' genes mapped to the network, (B): Non-pathogenic 'seed' genes mapped to the network, where (A) + (B) = (C): the total number of 'seed' genes mapped to the network, MPPI network – *Magnaporthe oryzae* protein-protein interaction network.

The data in Table 4-14 reveals that only a small number of candidate pathogenicity genes were predicted comparing to the number of 'seed' genes mapped to the network. This can result from either of MPPI network properties (Table 4-15) or the low mapping of *M. oryzae* genes from the genome assembly 6 to the interactome built with *M. oryzae* genes from the genome assembly 5.

4.4.2 Comparison of the MPPI network and the FPPI core network properties

The main properties and structure measurements such as average clustering coefficient, average degree centrality for the MPPI network were calculated and compared with the FPPI core network properties. The summary of this comparison is presented in Table 4-15.

Table 4-15 Comparison of the MPPI network with the FPPI core network.

Network	No of Nodes	No of Edges	No of connected components	Size of the largest connected component	Average Clustering Coefficient	Average Degree centrality
FPPI core network	2610	21690	260	1293	0.6266	0.0064
MPPI network	3016	11673	107	2892	0.1052	0.0026

By comparing the average clustering coefficient between MPPI and FPPI core networks, it can be observed that the genes within the FPPI core network are more likely to be connected to more genes within the network and act as hubs in the network. This is also visible in comparison of the relative number of edges to the number of nodes in both networks.

It is also interesting to note that the number of connected components in the MPPI network is relatively smaller when comparing to the FPPI network. Moreover, despite having relatively fewer edges number, the size of the largest connected component in the MPPI network accounts for 96% of all nodes in the MPPI network. This means that there is the potential for a greater spread of information within the MPPI network which appears to be more efficient compared to the predicted FPPI network.

4.4.3 Discussion

The analysis in this section has not been expanded further. This is because of the low number of predicted candidate genes in *M. oryzae* predicted PPI.

4.5 Conclusions

In this chapter the approach of Liu et al. (2010) was followed and the 'core network' (combined high confidence DDI and interologs networks) of Zhao et al. (2009) was used. In contrast to the study by Liu et al. (2010), careful dissection of experimentally verified *F. graminearum* genes into pathogenic ones and those that do not have an effect on the pathogenic process was performed.

On the basis of such classification, prediction of candidate genes responsible for pathogenicity and those that have no effect on the pathogenicity process in *F. graminearum* was possible. Moreover, each network dataset was examined individually calculating its main measurements such as average clustering coefficient, betweenness, and degree centrality. Additional analysis included characterisation of the *F. graminearum* genome with respect to MIPS functional categories of the genes, as well as sub-cellular location and identification of the position of the predicted pathogenic genes within the *F. graminearum* genome and available chromosomes map.

In summary, prediction of 65 potential candidate genes for pathogenicity in *F. graminearum* was performed. Most of those predicted candidate genes are thought to be a part of three main MAPK signalling cascades activated in *F. graminearum* during wheat ear infection. One of those predicted candidate genes, FGSG_06444 was highlighted to be a potential candidate for further biological experiments. This is because this gene was connected to the 'seed' gene FGSG_03537, which is the *TRI5* gene responsible for mycotoxins production and is specific to *Fusarium* genus.

Unfortunately, using the network approach described in this chapter it is not possible to predict all candidate genes for pathogenicity in *F. graminearum*. This is because there are several species-specific genes that are involved in trichothecene mycotoxins production and regulations, a process that is uniquely required for *F. graminearum* infection. Similarly, it was noted that there is a major underrepresentation in the FPPI of FG proteins predicted to reside outside the fungal cell. Further improvements to the study described in this chapter were made by Lysenko et al. (2013), in which I am a middle author (see Appendix F for an attached publication).

In this later study, Lysenko and Rothamsted colleagues employed a network analysis approach that combined the information from predicted FPPI network (Zhao et al., 2009), FG gene co-expression datasets from PLEXdb, and used FG sequence similarity to mapping 100 FG genes known to be vital for the virulence process (from PHI-base version 3.3, Jan 2012, see Table 2-1). By using this combined triple network, Lysenko et al. (2013) were able to predict 215 FG proteins to be potential candidates for pathogenicity. Among these predicted pathogenicity genes, 32 genes were also predicted to be a candidate for pathogenicity in my study (see Table 4-13 for highlighted in bold FG genes), where 20 of them were also predicted by Liu et al. (2010), leaving 12 FG predicted candidate genes in common for both my and Lysenko's study (Lysenko et al., 2013).

Chapter 5

Network approaches for exploring the role of Domains of Unknown Function (DUFs) in the disease-causing ability of the plant pathogenic fungus *Fusarium graminearum*

5.1 Introduction

Following an application of PPI network in the prediction of pathogenic genes in plant pathogenic fungi (Chapter 4), the study in this chapter concentrates mainly on the functional units of proteins, namely domains and their contributions to a protein functional annotation. The main attention in this study is focused on the Domains of Unknown Function (DUFs) present within the *F. graminearum* proteome. Proteins with DUFs and especially those with only domains that are DUFs are of interest in this study. These proteins are unlikely to be detected in PPI interolog networks due to the lack of significant sequence similarity to the annotated protein domain families available within public databases such as PFAM for example. On the other hand, DUFs might carry important information about the specific lifestyle of an organism of interest. Thus, the study in this chapter concentrates on DUFs in FG and their possible role in the disease-causing ability of this fungal plant pathogen.

Previously predicted candidate pathogenic genes in FG (Lysenko et al., 2013) together with the information from PHI-base (versions 3.4 to 3.6) are used to explore the role of DUFs in FG pathogenicity. Firstly, the pfam domain repertoire of the FG proteome is investigated with the main emphasis placed on the DUFs, their abundance within the proteome, and the location of the encoding genes within the four chromosomes. Moreover, a taxonomic diversity evaluation of pfam domains and DUFs is carried out for FG. Then, distinct domain-pair combinations (bigrams) are identified within the FG proteome. The bigrams are used further in the network analysis to examine the properties of DUFs and their possible impact on the pathogenic nature of FG.

5.2 Aims and objectives

The broad aims of the work in this chapter are

- Investigate the role of DUFs in the pathogenicity of the plant pathogenic fungus *Fusarium graminearum;*
- Identify DUFs repertoire specific to fungal species and their lifestyle;
- Provide metrics for DUFs and *F. graminearum* proteins function prediction.

This will be achieved through:

- Exploring pfam domain repertoire of *F. graminearum* proteome with the main emphasis placed on DUFs and their abundance within the proteome and genomic map of *F. graminearum.*
- Identifying distinct domain combinations (bigrams) within *F. graminearum* proteome concentrating mainly on *F. graminearum* proteins comprising of DUFs bigram(s).
- Taxonomic diversity evaluation of pfam domains and DUFs present in *F. graminearum.*
- Implementing the network analysis approach to examine the properties of DUFs and their possible impact on the pathogenic nature of *F. graminearum*.

5.3 Resources and methods

5.3.1 Identification of domain composition in *Fusarium graminearum*

Protein sequences were downloaded from MIPS *Fusarium graminearum* on 12/02/2013. The domain repertoire of FG proteins was identified using HMMER (biosequence analysis using profile hidden Markov models) algorithm implemented on TimeLogic® HMM (Hidden Markov Models) version 8.5.2.0 and domain models from Pfam_A database (version 27.0) (Finn et al., 2014a). For each protein with one or more pfam domain, all pfam domains were considered. This included situations where the same domain appears more than once in the given protein.

Potentially, any number of distinct domain signatures can be matched to the same region of the protein sequence. Therefore, additional processing of the raw HMMER output was necessary to resolve this ambiguity and generate a non-redundant representative set of domain bigrams. This was done by developing a customised computational pipeline in the Python programming language (solving_domains_overlapping.py, see https://github.com/ejsejda/PhD_thesis-Chapter_5/).

The general rules for solving the domain overlapping problem were adopted from previous work (Seidl et al., 2011) (see Figure 5-1). The rules were applied to compare the quality of each prediction in the overlapping set of predicted domains and decide which one was to be retained in a non-redundant set. In addition to this, a scoring system for each domain was introduced. Thus, if a domain did not overlap with the one compared to in the given protein, this domain was assigned a score equals to -1 and the domain remained in the protein.

To efficiently resolve the potentially complex situations where multiple domains overlapped, the set of overlapping domains was represented as an adjacency matrix, where the scores were assigned based on the application of the rules. To be specific, 1 was assigned to the row of the predicted domain if the rules indicated this domain was better, compared to the domain in the column, and 0 if the situation was the other way around. The domain with the score equal to 1 remained in the protein, whereas the domain with the score equal to 0 was removed from the protein sequence. The rules to solve the overlapping are shown in Figure 5-2. Although this approach resolved the overlap in most cases, there were 12 FG proteins where the overlapping had to be resolved manually and the justification for the choice is provided in Appendix D.

5.3.2 Domains bigrams analysis in *F. graminearum*

The domain bigram definition was adopted from a previous study (Seidl et al., 2011) as two successively located domains in a given protein. In this analysis domain repertoire of FG protein sequences and not sequence similarity between these proteins is studied. Thus, the order of domains with respect to N/C-terminus was not considered to be important. Therefore, bigram AB is the same as bigram BA and is defined as ‘hetero-bigram’ regarding its content. Repeated domains were also considered in the analysis – for example, neighbouring domains A and A

would count as an AA bigram and defined as 'homo-bigram' analogically. Figure 5-3 (adopted and modified figure from an earlier study (Seidl et al., 2011)) summarises the concept used in the bigrams identification within the predicted FG proteome. The frequency of each hetero-bigram and homo-bigram in the FG proteome was calculated with the aid of custom developed Python program (calculating_bigrams_statistic.py, https://github.com/ejsejda/PhD_thesis-Chapter_5/)

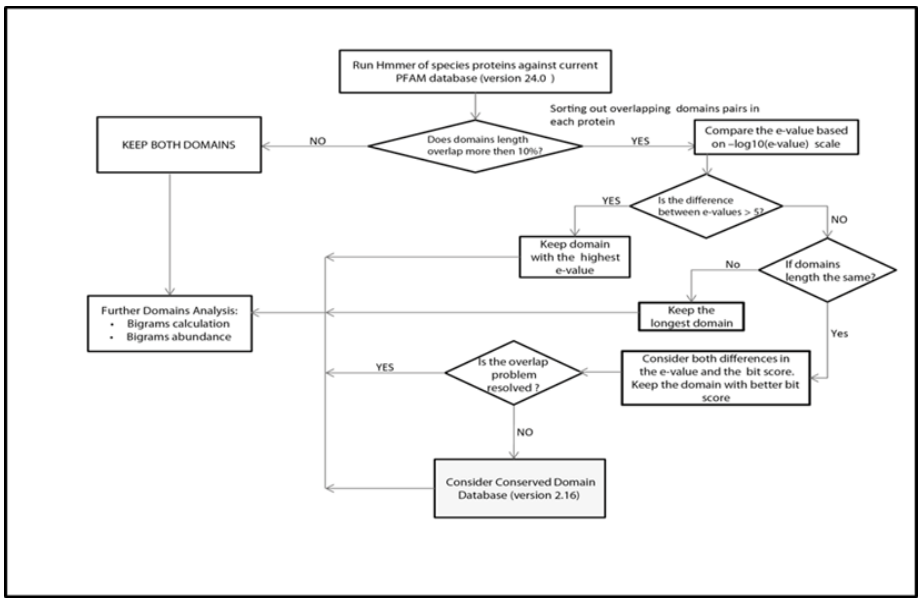


Figure 5-1 The flowchart to solve the domain overlapping issues in the study by Seidl et al. (2011).

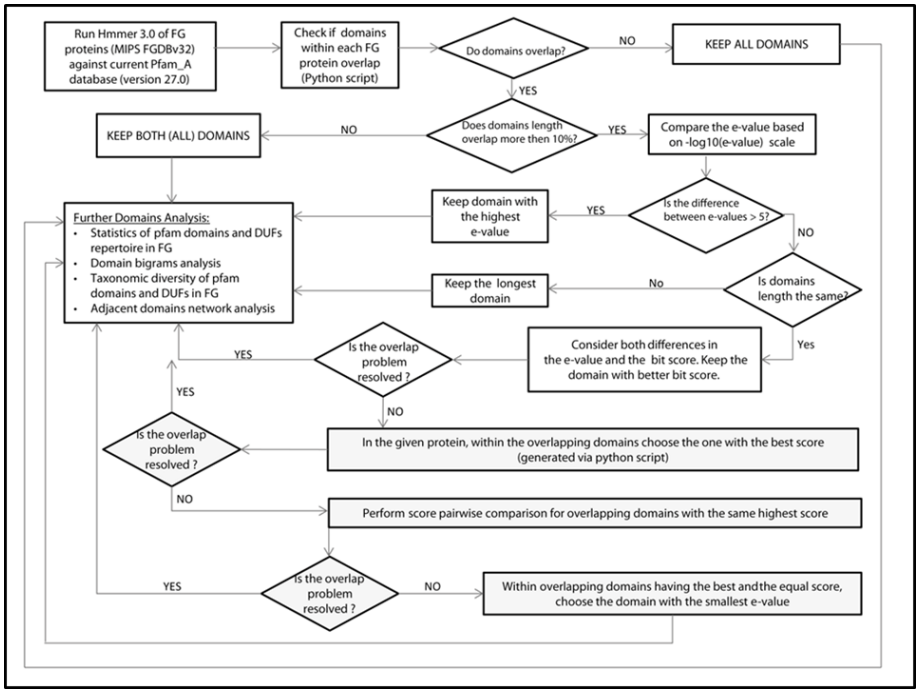


Figure 5-2 The flowchart to solve domain overlapping issues in *F. graminearum* proteome.

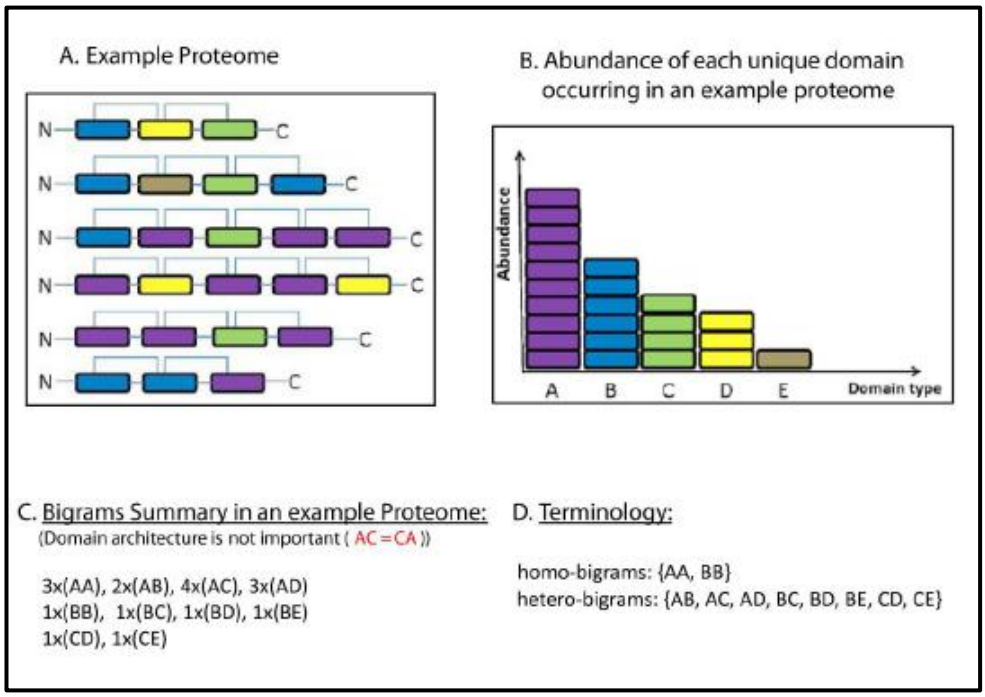


Figure 5-3 Metrics used for domain bigram analysis in *F. graminearum*.
Example proteome consists of five unique domains: A, B, C, D, and E.

5.3.3 Taxonomic diversity of pfam domains and DUFs identified within FG proteome

The information from PFAM database version 27.0, as well as UniProt database (Consortium, 2014) was used to evaluate taxonomic diversity of pfam domains as well as DUFs present in the FG proteome. Two custom-developed computational pipelines in the Python programming language were used to mine the necessary information from both resources: finding_taxaIds_for_pfam_domains.py and allDomainsTaxaInFG.py (see the github project: https://github.com/ejsejda/PhD_thesis-Chapter_5). The results were visualised using R software version 3.03. Statistical analysis of DUFs association with fungi lifestyles were also performed in R software version 3.03.

5.3.4 Network construction and analysis

The *Fusarium graminearum* domain-association network was generated using information delivered from pfam domains, including DUFs where applicable, that form 'hetero-bigrams' in FG proteome. The nodes of the network are pfam domains. The edges form between two domain nodes if these were part of at least one bigram. Each edge was assigned a normalised weight corresponding to the abundance of a given bigram within the whole FG.

The topological properties of the network, as well as the properties of each vertex (such as node degree, node clustering coefficient and node degree centrality) were calculated with NetworkX python package[32] (bigramNetwork.py, see https://github.com/ejsejda/PhD_thesis-Chapter_5/). Statistical tests comparing nodes properties were performed using R software version 3.03. The network was visualised using Cytoscape software, version 2.8.3.

5.3.5 Community structure detection

The community structure of the main connected component was identified by means of the greedy agglomerative algorithm knows as Louvain method (Vincent et al., 2008) (Figure 5-4). The method has the advantages of being very fast and accurate despite its greedy nature. The algorithm consists of two main phases: a modularity optimisation and a community aggregation.

In the first phase, each node represents an individual module. Then, successive iteration takes place over all nodes to verify which vertices should be connected to increase the modularity. The process is repeated until no further improvement in the modularity can be obtained. In the second phase of the algorithm, a new network is formed where the communities that have been established in the first phase become nodes in the new network and the links between those nodes are given the weight which is a sum of the weights of the links that join the two corresponding communities. Also, links between nodes in the same community become self-loops for this community in the new network. The step is repeated until no further gain in the modularity is achievable. As a result, it is possible to attain the best partitions of the initial network into communities.

Thus, looking at the network example depicted in Figure 5-4 (Vincent et al., 2008), 13 nodes (illustrated in light blue colour) represent 13 communities in the initial network. After modular optimisation and community aggregation, a new network of four nodes (green, blue, red, and light blue) is created. Then, both phases of the algorithm are repeated on the created network of four nodes. As consequence of the second pass of the algorithm, green and blue nodes fall into one community. Similarly, the red and light blue nodes becoming part of the second community of the

[32] http://networkx.github.io/documentation/latest/

third network. This is because a weight on the link between green and blue node is higher (4) comparing to a weight on the link between green and other two nodes (1), just as a weight on the link between red and light blue node (3) comparing to a weight on the link between red and green nodes (1). Thus, the blue node becomes a part of green node community and red node becomes a part of light blue node community leading to a weighted link between newly created communities equal to 3 (summary of links between green and light blue nodes, green and red nodes, and blue and light blue nodes). Finally, no further improvement in the modularity of a newly created third network could be obtained.

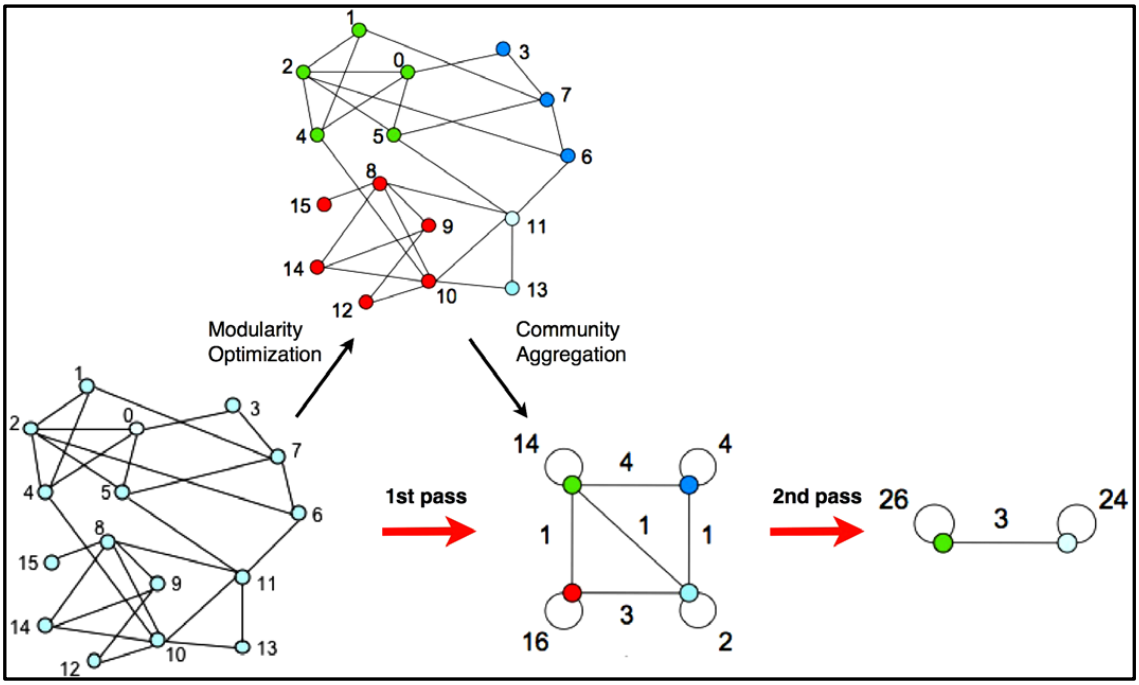


Figure 5-4 Visualisation of the steps of Louvain algorithm.
Where red, blue, green and light blue nodes indicate four different communities (modules). Weights of links between new nodes are the sum of the weight of the links between nodes in the corresponding two communities. Weight on new nodes is the number of self-loops calculated as links between nodes in the same community. For example links for the blue community: L={(3,7), (7,3), (7,6), (6,7)} account for 4 self-loops after 1st pass of the algorithm (Vincent et al., 2008).

5.3.6 Role of the domain nodes in the domain-association network

A node role is characterised according to two measures adopted from the Guimera et al. (2005) study: within-community degree z-score and participation coefficient P. Degree z-score measures the connectivity of the node to members of the same module, whereas participation coefficient likewise measures its connectivity to members of other modules relative to its own module. The

high value of a z-score indicates the high within-cluster node degree. The participation coefficient of a node is close to 1 if the links from a node are equally distributed among all clusters and is equal to 0 if all links of a node are within its own cluster.

The node classification scheme applied in this work was defined previously (Guimera and Nunes Amaral, 2005) and can be summarised as follows. Based on the region in a parameter space of z-score and participation coefficient, nodes are categorised as hubs (with a higher number of links within its own cluster) and non-hubs. Non-hubs nodes are further assigned four different roles: R1 - ultra-peripheral node (with all links within its cluster), R2 – peripheral node (with most links within its cluster), R3 - non-hub connector node (with many links to other clusters) and R4 – non-hub kinless node (with links homogeneously spread among all clusters). The hub nodes, however, are divided into further three categories: R5 – provincial hub (hub node with the great majority of links within its cluster), R6 – connector hub (hub with many links to other clusters) and R7 – global kinless hub (hub with links homogeneously spread among all clusters).

The roles of the nodes were determined using GIANT version 1.0 plugin for Cytoscape version 2.8.3.

5.3.7 Functional coherence and the community structure

The Average Information Content of the Most Informative Common Ancestor (AIC-MICA) metric (Lysenko et al., 2011) was used for identifying the most representative GO terms in the GO hierarchy that could best summarise the functional composition of a module. Using this method, a set of representative Most Informative Common Ancestor (MICA) terms can be identified, where the information content (IC) is computed based on the frequency of specific annotation found in an annotation set for a given species.

The AIC-MICA method takes as an input a set of annotated units and returns a non-redundant set of MICA terms that describe the specified proportion of the entities within a set. Average Information Content (AIC) associated with a set of MICA of certain coverage was calculated based on IC values. The higher value indicates that the most of MICAs for a tested module were found at a lower level in the ontology tree and would represent a functionality meaningful group. In

contrast, the lower value of AIC indicates that MICA would be close to the root of the ontology tree and would not represent a functionality meaningful group.

Here, the annotation for all three aspects of gene ontology for the modules detected by the Louvain method with at least five annotated nodes was considered and AIC-MICA approach was applied to find the most specific terms applicable to at least 30% of the nodes for a given module.

5.4 Results

5.4.1 Domain repertoire in *Fusarium graminearum* proteome

The abundance of each pfam domain within the predicted FG protein repertoire (n=13,826) was computed and the results are presented as a genome-wide distribution of pfam domains as Table 5-1 and Figure 5-5 where a part of the distribution is displayed. The predicted *F. graminearum* FG3 (MIPS) proteome, currently (at the time of writing the chapter) has 4,915 conserved hypothetical proteins and 3,034 hypothetical proteins with neither experimental nor computationally predicted functions assigned. In the PFAM version 27.0 there are 25% (3695) of DUF entries.

The first aim of this analysis was to identify DUFs within the FG proteome. Based on the information in pfamA.txt file from PFAM version 27.0, it was possible to rename PFAM Ids to the DUF naming scheme (where appropriate) and concentrate on DUFs presence within the FG proteome.

Overall, 61% (8,478) of all proteins in FG have one or more predicted pfam domain (Figure 5-5 and Table 5-1), leaving 39% (5,348) of the proteome being unannotated. Most FG proteins (43.55%), however, have only one pfam domain and in 5.6% (338) of these proteins the domain is a DUF.

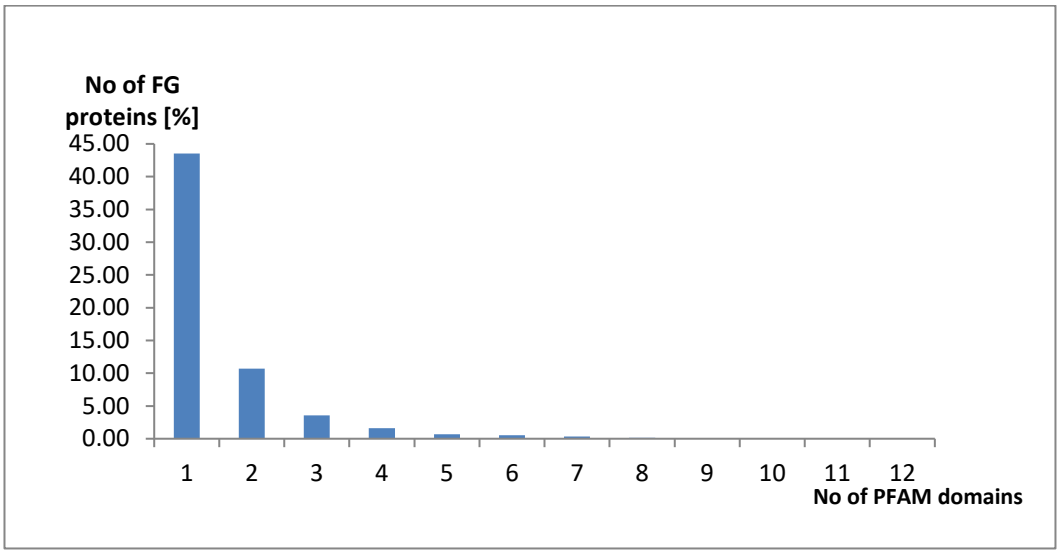


Figure 5-5 Distribution of pfam domains in the *F. graminearum* proteome encoded in FG3 MIPS genome assembly.

Table 5-1 PFAM domain distribution in the *F. graminearum* proteome.

No of PFAM domains per protein	FG proteins No	FG proteins No [%]
1	6021	43.55
2	1477	10.68
3	492	3.56
4	220	1.59
5	95	0.69
6	72	0.52
7	45	0.33
8	23	0.17
9	10	0.07
10	6	0.04
11	2	0.01
12	4	0.03
13	1	0.01
14	1	0.01
15	1	0.01
16	1	0.01
17	1	0.01
18	1	0.01
24	2	0.01
31	1	0.01
33	1	0.01
55	1	0.01

The chromosomal distribution of the genes coding for FG proteins with a single DUF domain within the entire FG genome is presented in Figure 5-6 generated with the aid of OmniMapFree (Antoniw et al., 2011).

In addition, the previously predicted proteins to be associated with the virulence and experimentally tested genes shown to be involved in the pathogenic lifestyle of FG (Lysenko et al., 2013) are also depicted in Figure 5-6. FG proteins with only one pfam domain that is a DUF are distributed evenly throughout the four chromosomes. None of these proteins was predicted in the study by Lysenko et al. (2013). One of the FG proteins with only one domain that is a DUF (DUF619), namely FGSG_01939, was experimentally proven to be required for virulence and was listed as a 'seed' gene in the previous study (Lysenko et al., 2013).

There are 23 FG proteins with 10 or more pfam domains within their sequence (Table 5-2). The table lists these domains together with the name of the encoding gene (where available). Most of these proteins either belong to or are related to non-ribosomal peptide synthetase (NRPS), while one multi-domain protein (FGSG_07798) is a probable polyketide synthase (PKS).

Both PKS and NRPS are multi-modular enzymes that have a characteristic modular structure and are encoded by a number of genes located in a cluster in the genome (Meier and Burkart, 2009, Strieker et al., 2010, Weissman and Leadlay, 2005). Although the modular structures of NRPSs and PKSs proteins are similar, the content of the modules are not identical (Appendix C, Figure C-1 and C-2). Several NRPSs and PKSs have been tested for functionality in various *Fusarium* species including *F. graminearum*. Their known products include the secreted metabolites fusarins, malonichrome or ferricrocin (products of PKS10, NRPS and NRPS2 respectively). These are bioactive secondary metabolites (mycotoxins) that lead to health problem if consumed by animals or humans and may play an important role in drug development or have an unknown function (Hansen et al., 2012).

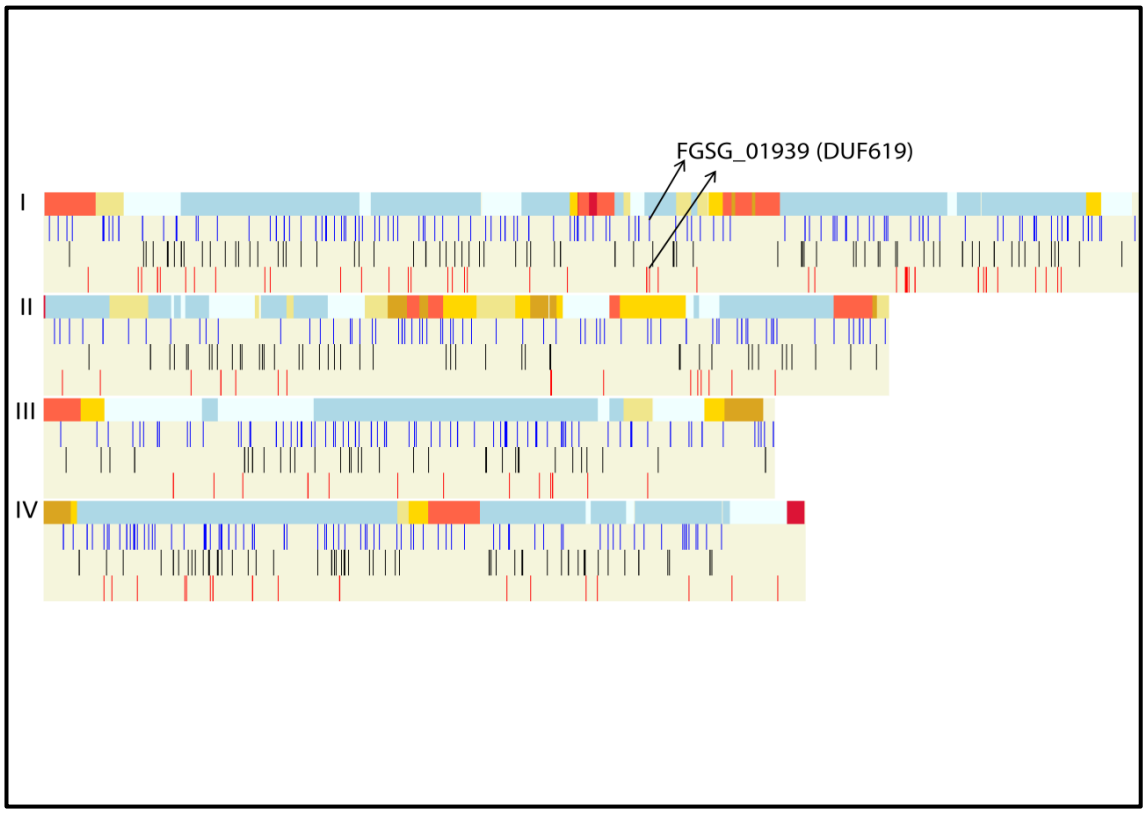


Figure 5-6 Position of *F. graminearum* genes coding for proteins with only one pfam domain that is a DUF along with other virulence-associated genes previously predicted (Lysenko et al., 2013) or experimentally verified virulence genes.
The diagram displays the four *F. graminearum* (FG) chromosomes indicated by Roman numerals. Recombination frequency across the chromosomes is depicted in track 1 using a colour gradient from white (0.0) lowest to crimson (>8 centimorgan (cM) highest). The various colours in the track 1 for each chromosome indicate the frequencies of the recombination (cM/27 kb), i. e. # clBeige 1 clKhaki 2 clGold 3 clGoldenRod 4 clTomato 8 clCrimson. The numbers between the colours are boundary values in cM/27 kb. Beige stands for the lowest, while crimson the highest recombination frequency. For each chromosome, blue coloured vertical bars in the track 2 locate the FG genes coding for proteins with only one pfam domain that is a DUF (n = 338). The predicted virulence genes from the previous study by Lysenko et al. (2013) are shown as black vertical bars in track 3 for each chromosome. The experimentally verified virulence 'seeds' (red bars) are depicted in track 4. The highlighted FGSG_01939 gene was previously predicted as virulence gene and it has one pfam domain that is DUF.

The majority of proteins listed in Table 5-2 consist of well-characterised domains. The exception here is the protein FGSG_01693, which consists of ten of the same domain of unknown function (DUF3659). As expected this protein belongs to the conserved hypothetical protein family of unknown function. DUF3659 also appears as only pfam domain in another FG conserved hypothetical protein, namely FGSG_01694.

Table 5-2 Details of the subset of *F. graminearum* proteins with the largest number of pfam domains.

No	FG_ID	Pfam domains No	Protein size [aa]	Chromosome	Previous alias	Gene	Description (Product of the gene)
1	FGSG_12118	55	3138	I	FG1:fgd117-10; FG1:fg13423		conserved hypothetical protein
2	FGSG_17487	33	11197	III	FGSG_13878	NPS5	related to non-ribosomal peptide synthetase
3	FGSG_17386	31	9539	I	FGSG_13783; FGSG_13784; FGSG_13785; FGSG_13786	NPS18	related to non-ribosomal peptide synthetase
4	FGSG_15673	24	7855	I	FGSG_11659; FGSG_11660; FGSG_17599	NPS8	non-ribosomal peptide synthetase
5	FGSG_02315	24	7639	I	FG1:fgd116-350; FG1:fg02315; FG1:fg12878	NPS4	related to non-ribosomal peptide synthetase
6	FGSG_05372	18	4841	III	FG1:fgd217-330; FG1:fg05372	NPS2	non-ribosomal peptide synthetase (Ferricrocin)
7	FGSG_10676	17	1016	I	FG1:fgd446-530; FG1:fg10676		conserved hypotetical protein
8	FGSG_01234	16	2279	I	FG1:fgd62-110; FG1:fgd62-120; FG1:fg01234; FG1:fg12289	MAC1	probable adenylate cyclase
9	FGSG_17132	15	1598	II	FGSG_13472; FGSG_13473		related to RSA4 - WD-repeat protein required for maturation and efficient intra-nuclear transport or pre-60S ribosomal subunits
10	FGSG_00916	14	1758	I	FG1:fgd41-90; FG1:fg00916		probable DNA-directed RNA polymerase II largest chain
11	FGSG_07798	13	3920	IV	FG1:fgd320-320; FG1:fg07798; FG1:fg12100	PKS10	probable polyketide synthase (***FUSS*** **, Fusarins**)
12	FGSG_11026	12	4747	III	FG1:fgd457-680; FG1:fg11026	NPS1	non-ribosomal peptide synthetase (**Malonichrome**)
13	FGSG_08956	12	1162	IV	FG1: fgd367-10; FG1: fg08956		related to kinesin light chain
14	FGSG_17133	12	1323	IV	FGSG_08952; FGSG_13475		related to RSA4 - WD-repeat protein required for maturation and efficient intra-nuclear transport or pre-60S ribosomal subunits
15	FGSG_08209	12	4423	II	FG1:fgd329-430 FG1:fg08209	NPS7	non-ribosomal peptide synthetase
16	FGSG_05619	11	1680	III	FG1: fgd231-60; FG1: fg05619		probable clathrin heavy chain
17	FGSG_07055	11	960	IV	FG1: fgd294-60; FG1: fg07055		probable DIP2 - Dom34p-interacting protein
18	FGSG_01693	10	1391	I	FG1: fgd91-20; FG1: fg01693		conserved hypothetical protein
19	FGSG_09638	10	2228	IV	FG1:fgd398-260 FG1:fg09638		probable URA2 - multifunctional pyrimidine biosynthesis protein
20	FGSG_16370	10	1693	II	FGSG_12558		hypothetical protein
21	FGSG_06788	10	907	IV	FG1: fgd275-470; FG1: fg06788		related to UTP13 - U3 snoRNP protein
22	FGSG_17092	10	1249	II	FGSG_08155		hypothetical protein
23	FGSG_15796	10	821	I	FGSG_00909		related to Sel-1 homolog precursor

Where aa stands for amino acids

In addition to analysing the distribution of pfam domains within the FG proteome, the overall domain abundance was calculated for each pfam domain present. In this calculation, all pfam domains that appear at least once in any FG protein were considered. In total 13,217 pfam domains were detected. This includes 3,524 unique domains. Domains of unknown function represent only 3.81% (504) of the total pfam domain repertoire. This accounts for 314 unique DUFs. The most abundant pfam domains are listed in Table 5-3. Further inspection of Table 5-3 reveals that the function of the most abundant pfam domains is well defined. Moreover, some of the most abundant pfam domains (PF00172 and PF04082) are fungal-specific domains (Fungal Zn(2)-Cys(6) binuclear cluster domain and fungal specific transcription factor domain respectively). As expected, the abundance of each unique DUF is low and varies from 16 to 1 occurrence.

The frequency of DUFs is presented in Figure 5-7. The most frequent DUFs are DUF3433 (16), DUF3425 (12), DUF3659 (11) and DUF2235 (8) (the number in brackets indicates the frequency of the domain as per Figure 5-7). The most frequent DUF3433 is present in 9 FG proteins: FGSG_00063, FGSG_04690, FGSG_05634, FGSG_10168, FGSG_16141, FGSG_16142, FGSG_16575, FGSG_16601 and FGSG_17008. The first two FG proteins listed above have only one domain which is DUF3433, while the rest of them consist of two copies of this DUF. On average, DUF3433 takes up less than 20% of the whole protein length when present.

The second most frequent domain of unknown function is DUF3425 which is present as a single copy in 11 different FG proteins with the average coverage of 35% of the whole protein length. Most DUF3425 domains exist without any other pfam domains in the protein sequence. The exception here is FGSG_12345 protein that consists of one copy of DUF3425 and one copy of PF00170 (bZIP transcription factor). However, PF00170 domain accounts for only 15% of the whole protein length here.

DUF3659 is also a frequent DUF in the proteome. Although this domain appears 11 times, its presence is only noted in two FG proteins: FGSG_01693 and FGSG_01694. There are ten copies of DUF3659 in the first protein, accounting for 46% of the whole protein length. Whereas, protein FGSG_01694 holds only one copy of DUF3659 domain and the domain accounts for 23% of the

protein length. It is interesting to note that DUF3659 is found in 17 different architectures within the PFAM version 27.0.

Table 5-3 Occurrence and function of the most abundant pfam domains within the *F. graminearum* proteome.

No	PFAM Id	Abundance in FG	Domain description
1	PF00400	464	WD domain, G-beta repeat
2	PF00172	301	Fungal Zn(2)-Cys(6) binuclear cluster domain
3	PF12796	251	Ankyrin repeats (3 copies)
4	PF07690	244	Major Facilitator Superfamily
5	PF04082	180	Fungal-specific transcription factor domain
6	PF13894	150	C2H2-type zinc finger
7	PF00153	117	Mitochondrial carrier protein
8	PF00005	111	ABC transporter
9	PF00083	110	Sugar (and other) transporter
10	PF00106	107	short chain dehydrogenase
11	PF00067	105	Cytochrome P450
12	PF00069	105	Protein kinase domain
13	PF00023	103	Ankyrin repeat
14	PF06985	103	Heterokaryon incompatibility protein (HET)
15	PF00501	86	AMP-binding enzyme
16	PF00076	81	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
17	PF00271	76	Helicase conserved C-terminal domain
18	PF00550	76	Phosphopantetheine attachment site
19	PF05792	74	Candida agglutinin-like (ALS)
20	PF00668	63	Condensation domain
21	PF12697	63	Alpha/beta hydrolase family
22	PF00004	61	ATPase family associated with various cellular activities (AAA)
23	PF13193	57	AMP-binding enzyme C-terminal domain
24	PF00107	57	Zinc-binding dehydrogenase
25	PF00664	55	ABC transporter transmembrane region
26	PF02178	53	AT hook motif
27	PF13489	53	Methyltransferase domain
28	PF05729	51	NACHT domain
29	PF08240	49	Alcohol dehydrogenase GroES-like domain
30	PF13374	48	Tetratricopeptide repeat
31	PF00270	47	DEAD/DEAH box helicase
32	PF01565	44	FAD binding domain
33	PF05368	42	NmrA-like family
34	PF00096	41	Zinc finger, C2H2 type
35	PF11951	41	Fungal-specific transcription factor domain
36	PF01494	38	FAD binding domain
37	PF13561	37	Enoyl-(Acyl carrier protein) reductase
38	PF08238	36	Sel1 repeat
39	PF01061	34	ABC-2 type transporter
40	PF13414	33	TPR repeat
41	PF01822	31	WSC domain

Moreover, these architectures include several copies of this domain. This might suggest that this domain is likely to be present in several copies in one protein. This is visible while examining the domain content of the FGSG_01693 protein. The question arises here, if proteins FGSG_01693 and FGSG_01694 should form one protein with the 11 copies of DUF3659. This is because protein FGSG_01694 is a relatively short protein (273 amino acids (aa)) with a one copy of DUF3659 and no other pfam domain is present. Both proteins are conserved hypothetical proteins and both reside on chromosome I within a region of high recombination frequency (Antoniw et al., 2011).

The fourth most frequent domain is DUF2235. This domain appears as a single copy in eight FG proteins and occupies on average 60% of the protein length and with no other pfam domain present within the protein.

All high-frequency DUFs are quite diverse in terms of sequence similarity and none of them appears in the FG proteins that have been tested experimentally for function.

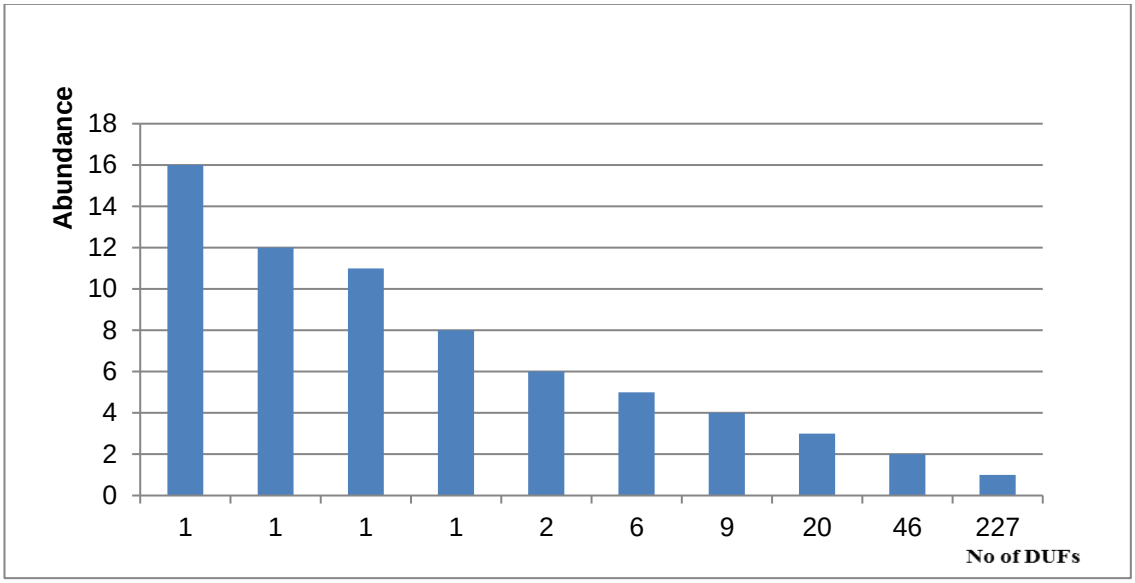


Figure 5-7 Frequency of DUFs in *F. graminearum* proteome.

5.4.2 Domain bigrams analysis in *Fusarium graminearum*

From the 2,457 FG proteins that contain two or more pfam domains (Table 5-1) available for this study, in total 4,739 bigrams were generated accounting for 1,687 unique bigrams within the FG genome. The majority of unique bigrams (1,523, 90.27%) are hetero bigrams. One of the homo-bigram, namely PF00400|PF00400 appears 343 times in the whole *F. graminearum* (FG) proteome but only in 96 FG proteins. This indicates that this bigram appears more than once in certain FG proteins. With regards to hetero-bigrams, bigram PF00172|PF04082 is highly represented in FG proteome and appears in 121 different proteins (Table 5-4).

As expected (Table 5-4), the most frequent bigrams occurred in FG proteome consisted of four of the most widespread pfam domains within the proteome: PF00400|PF00400, PF12796|PF12796, and PF00172|PF04082 (Table 5-3). The first two represent the most frequent homo-bigrams, whereas the third one is a hetero-bigram. Both domains in the hetero-bigram are fungal-specific and are found within Transcription Associated Proteins (TAPs).

Overall, in total 153 (3.23% of total bigram generated) bigrams with at least one DUF were generated. This includes 108 unique bigrams with at least one DUF domain listed in Table 5-5. Out of these 153 bigrams, 40 bigrams have only DUF pairs which account for 20 unique bigrams with both DUFs which are listed in Table 5-6.

The frequency of domain bigrams consisting of at least one DUF domain is very low and varies in occurrence from nine to one within the entire FG genome. Two the most abundant bigrams are homo-bigrams: DUF3659|DUF3659 and DUF3433|DUF3433. The first one appears only in one protein (FGSG_01693), while the second one is present in seven different FG proteins. None of the FG proteins with the most abundant DUF bigrams has been experimentally verified as required for pathogenicity.

Table 5-4 The most frequent hetero-bigrams and homo-bigrams within *F. graminearum* proteome.

No	Hetero - bigrams					Homo - bigrams				
	Domains bigrams		Total No in FG proteome (N)	Unique No in FG proteome	Weight (W) $W = N/N_{TOTAL}$ $N_{TOTAL} = 121$	Domains bigrams		Total No in FG proteome (N)	Unique No in FG proteome	Weight (W) $W = N/N_{TOTAL}$ $N_{TOTAL} = 343$
1	PF00172	PF04082	121	121	1.00	PF00400	PF00400	343	96	1.00
2	PF00005	PF00664	81	32	0.67	PF12796	PF12796	151	67	0.44
3	PF00501	PF13193	56	38	0.46	PF13894	PF13894	78	49	0.23
4	PF00550	PF00668	53	12	0.44	PF00153	PF00153	75	38	0.22
5	PF00270	PF00271	45	44	0.37	PF05792	PF05792	71	3	0.21
6	PF00107	PF08240	39	39	0.32	PF00023	PF00023	67	18	0.20
7	PF00501	PF00668	35	10	0.29	PF13374	PF13374	40	7	0.12
8	PF00005	PF01061	35	19	0.29	PF02178	PF02178	34	16	0.10
9	PF00550	PF13193	32	14	0.26	PF08238	PF08238	30	6	0.09
10	PF00172	PF11951	30	30	0.25	PF00076	PF00076	29	18	0.08
11	PF00122	PF00702	24	24	0.20	PF00098	PF00098	20	4	0.06
12	PF00096	PF13894	23	22	0.19	PF00806	PF00806	16	4	0.05
13	PF00732	PF05199	21	21	0.17	PF00013	PF00013	15	5	0.04
14	PF00176	PF00271	20	20	0.17	PF00514	PF00514	14	2	0.04
15	PF00933	PF01915	19	19	0.16	PF00515	PF00515	14	4	0.04
16	PF00501	PF00550	19	15	0.16	PF00004	PF00004	13	11	0.04
17	PF00109	PF02801	18	18	0.15	PF01822	PF01822	13	6	0.04
18	PF01565	PF08031	18	18	0.15	PF00668	PF00668	13	4	0.04
19	PF00023	PF12796	18	12	0.15	PF02985	PF02985	12	4	0.03
20	PF01915	PF14310	18	18	0.15	PF00432	PF00432	11	4	0.03
21	PF01794	PF08022	17	17	0.14	PF13415	PF13415	10	5	0.03
22	PF00512	PF02518	16	16	0.13	PF00415	PF00415	9	4	0.03
23	PF00072	PF02518	16	16	0.13	PF13516	PF13516	9	6	0.03
24	PF05729	PF12796	16	16	0.13	PF01699	PF01699	9	9	0.03
25	PF01061	PF06422	15	15	0.12	PF02012	PF02012	9	3	0.03
26	PF00698	PF02801	15	15	0.12	DUF3659	DUF3659	9	1	0.03
27	PF08022	PF08030	15	15	0.12	PF05001	PF05001	8	1	0.02
28	PF00122	PF00690	15	15	0.12	PF13465	PF13465	8	4	0.02
29	PF00005	PF06422	14	14	0.12	PF00571	PF00571	8	4	0.02
30	PF00394	PF07732	14	14	0.12	PF00096	PF00096	7	6	0.02
31	PF00394	PF07731	14	14	0.12	PF00560	PF00560	7	3	0.02
32	PF00698	PF14765	14	14	0.12	PF00575	PF00575	7	1	0.02
33	PF00009	PF03144	13	13	0.11	PF00168	PF00168	7	4	0.02
34	PF00689	PF00702	13	13	0.11	PF01476	PF01476	7	4	0.02
35	PF13086	PF13087	12	12	0.10	PF04193	PF04193	7	7	0.02
36	PF00515	PF13414	12	6	0.10	DUF3433	DUF3433	7	7	0.02
37	PF00005	PF14510	12	12	0.10	PF01239	PF01239	7	2	0.02
38	PF00082	PF05922	11	11	0.09	PF10281	PF10281	6	1	0.02
39	PF00043	PF13417	11	11	0.09	PF00612	PF00612	6	2	0.02
40	PF00441	PF02770	11	11	0.09	PF00637	PF00637	6	1	0.02
41	PF00175	PF00970	11	11	0.09	PF02469	PF02469	6	6	0.02
42	PF02225	PF04389	10	10	0.08	PF13517	PF13517	6	2	0.02
43	PF00550	PF07993	10	10	0.08	PF13634	PF13634	6	3	0.02
44	PF03171	PF14226	9	9	0.07	PF07719	PF07719	6	4	0.02
45	PF04408	PF07717	9	9	0.07	PF07728	PF07728	6	2	0.02

Where: Unique No in *F. graminearum* proteome indicates the number of proteins the bigram appears in; Weight (W) – is a normalised frequency of the total bigrams number and is equal to 1 for the most frequent hetero - bigram and the most frequent homo - bigram within the *F. graminearum* proteome. The bigram PF00172|PF04082 is the most frequent bigram in this analysis and it appears 121 times within the *F. graminearum* proteome.

Table 5-5 Frequency of bigrams with at least one DUF in the *F. graminearum* proteome.

No	Bigram	No in FG*	FG No**	No	Bigram	No in FG*	FG No**	No	Bigram	No in FG*	FG No**
1	DUF3659|DUF3659	9	1	37	DUF3395|PF00226	1	1	73	DUF4139|DUF4140	1	1
2	DUF3433|DUF3433	7	7	38	DUF500|PF07653	1	1	74	DUF1981|PF01369	1	1
3	DUF4463|PF13967	5	5	39	DUF4414|DUF913	1	1	75	DUF3818|PF00787	1	1
4	DUF4217|PF00271	4	4	40	DUF1214|DUF1254	1	1	76	DUF3694|PF00169	1	1
5	DUF221|DUF4463	4	4	41	DUF3543|PF00069	1	1	77	DUF3517|PF00443	1	1
6	DUF3336|PF01734	3	3	42	DUF908|DUF913	1	1	78	DUF2405|PF10344	1	1
7	DUF2422|PF13515	3	3	43	DUF1227|PF06733	1	1	79	DUF1929|PF07646	1	1
8	DUF3638|DUF3645	2	2	44	DUF1227|PF13307	1	1	80	DUF1932|PF03446	1	1
9	DUF3535|PF02985	2	1	45	DUF4110|PF13415	1	1	81	DUF4210|PF13889	1	1
10	DUF221|DUF3779	2	2	46	DUF3419|PF13489	1	1	82	DUF1998|PF00271	1	1
11	DUF1929|PF01344	2	2	47	DUF2411|DUF2435	1	1	83	DUF3608|PF00610	1	1
12	DUF917|PF01968	2	2	48	DUF382|PF04046	1	1	84	DUF1162|PF09333	1	1
13	DUF1720|PF12763	2	1	49	DUF3735|PF12430	1	1	85	DUF367|PF04068	1	1
14	DUF4470|PF01753	2	2	50	DUF3546|DUF4187	1	1	86	DUF3639|PF00400	1	1
15	DUF1729|PF13452	2	2	51	DUF3554|PF02985	1	1	87	DUF1929|PF07250	1	1
16	DUF1965|PF01179	2	2	52	DUF619|PF00696	1	1	88	DUF1620|PF13360	1	1
17	DUF1771|PF01713	2	2	53	DUF3694|PF12423	1	1	89	DUF1996|PF01822	1	1
18	DUF1212|DUF3815	2	2	54	DUF1982|PF00384	1	1	90	DUF3402|PF07923	1	1
19	DUF2421|PF13515	2	2	55	DUF1900|PF00400	1	1	91	DUF383|DUF384	1	1
20	DUF202|PF09359	2	2	56	DUF1115|PF08572	1	1	92	DUF2401|DUF2403	1	1
21	DUF2427|PF10355	2	2	57	DUF1771|PF00642	1	1	93	DUF3471|PF00144	1	1
22	DUF663|PF08142	2	2	58	DUF3506|PF00646	1	1	94	DUF3385|PF02985	1	1
23	DUF1785|PF02170	2	2	59	DUF1720|PF01417	1	1	95	DUF1899|PF00400	1	1
24	DUF1034|PF00082	2	2	60	DUF1726|PF05127	1	1	96	DUF1744|PF00136	1	1
25	DUF1446|DUF4387	1	1	61	DUF1546|PF02969	1	1	97	DUF4414|PF00632	1	1
26	DUF1929|PF13418	1	1	62	DUF2156|PF00152	1	1	98	DUF1720|PF03983	1	1
27	DUF3441|PF05833	1	1	63	DUF3449|PF12171	1	1	99	DUF1965|PF02727	1	1
28	DUF1752|PF00320	1	1	64	DUF3384|PF03542	1	1	100	DUF2405|PF10305	1	1
29	DUF3385|PF02259	1	1	65	DUF3819|PF04054	1	1	101	DUF3835|PF13758	1	1
30	DUF307|PF01699	1	1	66	DUF1977|PF00226	1	1	102	DUF4208|PF00271	1	1
31	DUF126|DUF521	1	1	67	DUF2014|PF00010	1	1	103	DUF1691|DUF1691	1	1
32	DUF504|PF04926	1	1	68	DUF1162|PF12624	1	1	104	DUF3818|PF12828	1	1
33	DUF3814|PF01262	1	1	69	DUF2347|DUF4484	1	1	105	DUF4187|PF01585	1	1
34	DUF1752|DUF3295	1	1	70	DUF3381|PF07780	1	1	106	DUF1604|PF01585	1	1
35	DUF2407|DUF2407_C	1	1	71	DUF21|PF00571	1	1	107	DUF3381|PF01728	1	1
36	DUF3337|PF00400	1	1	72	DUF3814|PF02233	1	1	108	DUF3425|PF00170	1	1

*Total occurrence of bigram in FG proteome

** Number of FG proteins the bigram appears in.

Table 5-6 Frequency of DUF bigrams in the *F. graminearum* proteome.

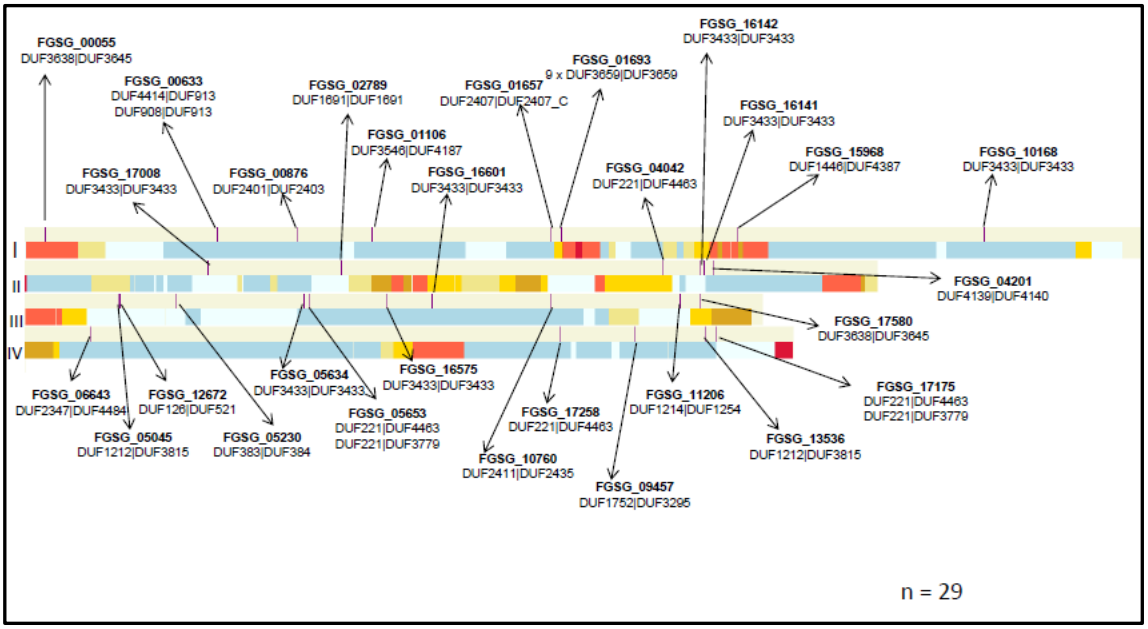
Bigram	Corresponding pfam ID	No in FG *	FG No **
DUF3659|DUF3659	PF12396|PF12396	9	1
DUF3433|DUF3433	PF11915|PF11915	7	7
DUF221|DUF4463	PF02714|PF14703	4	4
DUF3638|DUF3645	PF12340|PF12359	2	2
DUF221|DUF3779	PF02714|PF12621	2	2
DUF1212|DUF3815	PF06738|PF12821	2	2
DUF1446|DUF4387	PF07287|PF14330	1	1
DUF126|DUF521	PF01989|PF04412	1	1
DUF1752|DUF3295	PF08550|PF11702	1	1
DUF2407|DUF2407_C	PF10302|PF13373	1	1
DUF4414|DUF913	PF14377|PF06025	1	1
DUF1214|DUF1254	PF06742|PF06863	1	1
DUF908|DUF913	PF06012|PF06025	1	1
DUF2411|DUF2435	PF10304|PF10363	1	1
DUF3546|DUF4187	PF12066|PF13821	1	1
DUF2347|DUF4484	PF09804|PF14831	1	1
DUF4139|DUF4140	PF13598|PF13600	1	1
DUF383|DUF384	PF04063|PF04064	1	1
DUF2401|DUF2403	PF10287|PF10290	1	1
DUF1691|DUF1691	PF07950|PF07950	1	1

Where FG – *F. graminearum*; *Total occurrence of bigram in FG proteome;
** Number of FG proteins the bigram appears in.

However, one of the FG proteins highlighted in Figure 5-8 B (generated using OmniMapFree (Antoniw et al., 2011)), namely FGSG_01106, has been experimentally verified as required for virulence. This protein consists of only two pfam domains which both are DUF and forms a bigram DUF3546|DUF4187. This finding provides the direct evidence which might suggest that DUFs can be an important factor in the plant pathogen lifestyle, as observed in the previous study (Seidl et al., 2011)

Altogether, there are 29 FG proteins which consist of only DUF bigrams. Only 4 of them are within the highest recombination regions throughout the FG genome indicating that the majority of DUF enriched proteins are well conserved proteins. However, the most DUF-rich protein namely FGSG_01693 is present within the highest recombination region of FG genome (Figure 5-8, generated using OmniMapFree (Antoniw et al., 2011)). As expected, most FG proteins with DUF bigrams only are either conserved hypothetical proteins or hypothetical proteins (Table 5-7).

A)



B)

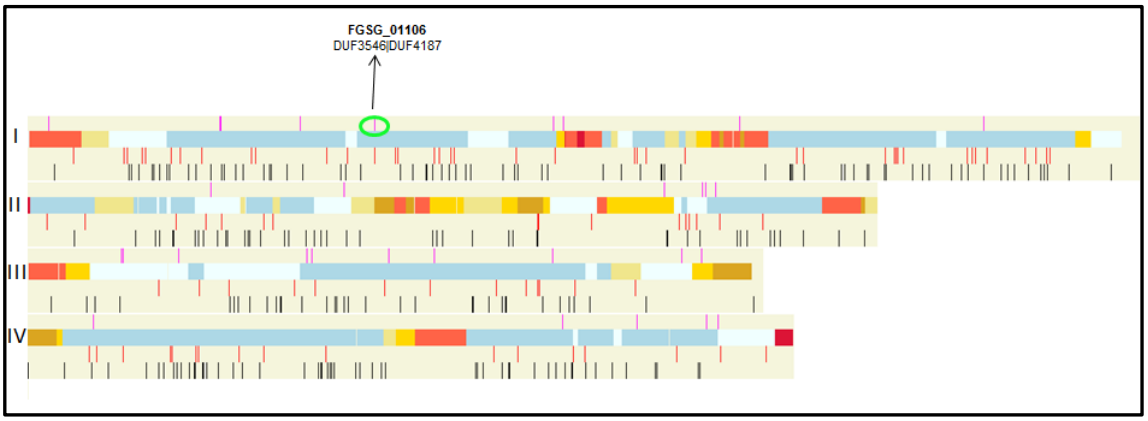


Figure 5-8 Position of genes coding for *F. graminearum* proteins with DUF bigrams along with other genes predicted by a previous network study or experimentally verified virulence genes.
The diagram in A) and B) displays the four *F. graminearum* (FG) chromosomes indicated by Roman numerals. Fuchsia vertical bars in the track 1 of each chromosome disclose FG proteins with DUF bigrams only. Recombination frequency across the chromosomes is depicted in track 2 using a colour gradient (white (0.0) lowest to crimson (>8 centimorgan (cM)) highest. The various colours in track 2 for each chromosome indicate the frequency of recombination (cM/27 kb), i. e. # clBeige 1 clKhaki 2 clGold 3 clGoldenRod 4 clTomato 8 clCrimson. The numbers between the colours are boundary values in cM/27 kb. Beige represents the lowest, whereas crimson the highest recombination frequency.
B) The experimentally verified virulence seeds (red bars) are depicted in track 3. The predicted virulence genes by a previous study (Lysenko et al., 2013) are shown as black vertical bars in track 4 for each chromosome.

Table 5-7 *F. graminearum* proteins with DUF bigrams only.

FGSG Id	Protein Description *	DUF bigram	WoLF PSORT subcellular localization prediction
FGSG_00055	conserved hypothetical protein	DUF3638|DUF3645	extr: 16, nucl: 5.5, cyto_nucl: 4.833, cyto_mito: 3.333, cyto: 3
FGSG_00633	related to TOM1 protein	DUF4414|DUF913, DUF908|DUF913	nucl: 12, cyto_nucl: 10, cyto: 6, mito: 4, plas: 4
FGSG_00876	conserved hypothetical protein	DUF2401|DUF2403	extr: 14, mito: 7, nucl: 2, vacu: 2
FGSG_01106	related to arsenite-resistance protein 2	DUF3546|DUF4187	nucl: 21.5, cyto_nucl: 13.5, cyto: 4.5
FGSG_01657	conserved hypothetical protein	DUF2407|DUF2407_C	plas: 20, nucl: 2, mito: 1, cyto: 1, E.R.: 1
FGSG_01693	conserved hypothetical protein	DUF3659|DUF3659 x 9	cyto: 14, cyto_nucl: 9.5, mito: 5, nucl: 3, pero: 2
FGSG_02789	conserved hypothetical protein	DUF1691|DUF1691	plas: 22, mito: 2, E.R.: 2
FGSG_04042	conserved hypothetical protein	DUF221|DUF4463	plas: 22, E.R.: 3
FGSG_04201	conserved hypothetical protein	DUF4139|DUF4140	nucl: 20, cyto_nucl: 13.833, mito_nucl: 10.833, cyto: 6.5
FGSG_05045	conserved hypothetical protein	DUF1212|DUF3815	plas: 22, nucl: 2, cyto: 1
FGSG_05230	conserved hypothetical protein	DUF383|DUF384	nucl: 15.5, cyto_nucl: 14.333, cyto: 11, mito_nucl: 8.666
FGSG_05634	conserved hypothetical protein	DUF3433|DUF3433	plas: 25
FGSG_05653	related to A.thaliana hyp1 protein	DUF221|DUF4463, DUF221|DUF3779	plas: 26
FGSG_06643	conserved hypothetical protein	DUF2347|DUF4448	mito: 14, nucl: 9.5, cyto_nucl: 6.5, cyto: 2.5
FGSG_09457	conserved hypothetical protein	DUF1752|DUF3295	nucl: 21, mito: 4
FGSG_10168	conserved hypothetical protein	DUF3433|DUF3433	plas: 23, E.R.: 2
FGSG_10760	conserved hypothetical protein	DUF2411|DUF2435	plas: 9, nucl: 7.5, cyto_nucl: 7.333, cyto: 6, cyto_mito: 4.333, vacu: 2
FGSG_11206	conserved hypothetical protein	DUF1214|DUF1254	extr: 25
FGSG_12672	conserved hypothetical protein	DUF126|DUF521	cyto_mito: 12.333, cyto: 12, mito: 11.5, cyto_nucl: 6.833
FGSG_13536	conserved hypothetical protein	DUF1212|DUF3815	plas: 23, nucl: 1, cyto: 1
FGSG_15968	hypothetical protein	DUF1446|DUF4387	cysk: 22, cyto_nucl: 3.5, cyto: 3
FGSG_16141	hypothetical protein	DUF3433|DUF3433	plas: 25
FGSG_16142	hypothetical protein	DUF3433|DUF3433	plas: 25
FGSG_16575	hypothetical protein	DUF3433|DUF3433	plas: 27
FGSG_16601	hypothetical protein	DUF3433|DUF3433	plas: 24, E.R.: 2
FGSG_17008	hypothetical protein	DUF3433|DUF3433	plas: 25
FGSG_17175	related to RSN1 - Overexpression rescues sro7/sop1 in NaCl	DUF221|DUF4463, DUF221|DUF3779	plas: 18, E.R.: 7
FGSG_17258	hypothetical protein	DUF221|DUF4463	plas: 23, vacu: 2
FGSG_17580	hypothetical protein	DUF3638|DUF3645	nucl: 19.5, cyto_nucl: 12.5, cyto: 4.5

*MIPS annotation

Where cellular compartments are cyto – cytosol, cysk – cytoskeleton, E.R.- endoplasmic reticulum, extr - extracellular location, plas - plasma membrane, mito – mitochondria, nucl – nucleus, vacu – vacuole.

Furthermore, WoLF PSORT[33] subcellular localisation prediction (with default setting: kNN = 27) of FG proteins with only DUF bigrams was performed and the results are listed in Table 5-7. As mentioned earlier (Chapter 4, section 4.3.2.2), WoLF PSORT is a program for a protein subcellular localisation prediction. It converts amino acids sequences into numerical vectors (numerical cell localisation), which are classified by simple k-nearest neighbour classifier. WoLF PSORT organises proteins into more than 10 localisation sites including dual localisation for proteins moving between cytosol and nucleus (Horton et al., 2007).

The analysis indicates that the majority of FG proteins with DUF only bigrams are intracellular proteins. Only three proteins namely FGSG_00055, FGSG_00876, and FGSG_11206 were identified to be in an extracellular compartment with the only FGSG_11206 showing strong evidence to be in the extracellular location. As mentioned previously in this chapter, the only one FG protein with both DUFs in a bigram, namely FGSG_01106, has been experimentally verified as required for virulence. The protein identified to be an intracellular protein with strong evidence to be in a nucleus.

5.4.3 Taxonomic diversity of pfam domains and DUFs in *F. graminearum*

In this analysis, based on the information from UniProt taxonomy34 and PFAM version 27.0, taxonomic diversity evaluation was performed for all domains (including DUF) present in FG. Next, the taxonomic diversity of only the DUFs identified in FG was analysed. The study focuses on the Prokaryotic kingdom Bacteria and three major Eukaryotic kingdoms: Fungi, Animals, and Plants, as well as other Eukaryotes including Oomycetes (Figure 5-9).

Overall, 3304 unique pfam domains of FG proteome, including domains labelled as DUF, were classified into the above taxonomic groups. This accounts for nearly 94% of all unique domains (3524) identified within FG proteome. It is still unclear why 6% of these domains was still unaccounted for. However, this discrepancy is most likely to be a result of the different version of PFAM and/or domain identification algorithm used to populate the UniProt database.

---

[33] http://www.genscript.com/psort/wolf_psort.html
[34] http://www.uniprot.org/taxonomy/

Pfam domains present in FG were also identified in a further 412 fungal species. Out of these fungal species, 218 have at least one DUF in common with FG. There are 155 Ascomycota, 45 Basidiomycota, and 18 other fungal species that share at least one DUF with FG. This occurrence pattern is expected because FG resides within the Ascomycota (see Chapter 2).

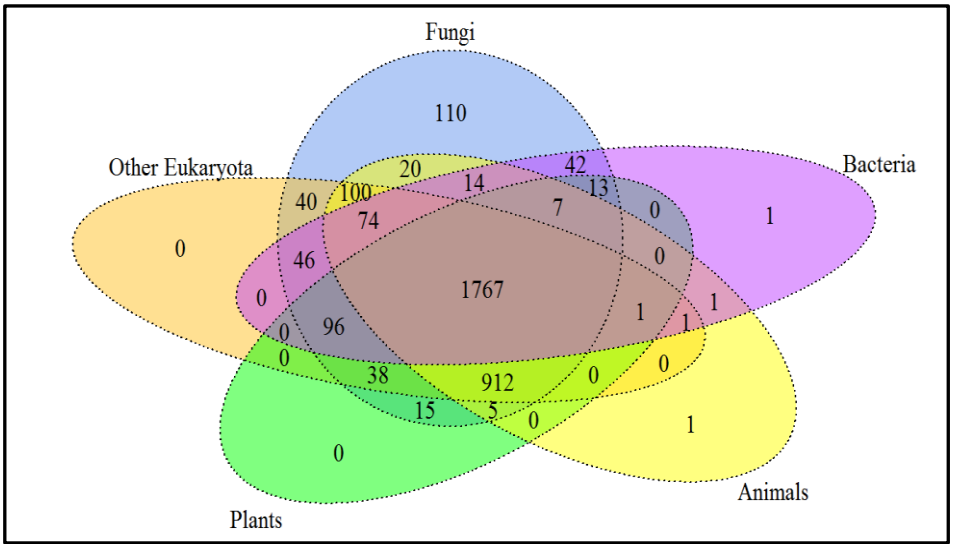
From the taxonomically classified domains, in total 110 were found to be fungi-specific. Furthermore, 35 of the fungi-specific domains were DUFs (Figure 5-9). These were found in 210 fungal species including 55 plant fungal pathogens together with FG. Five pfam domains were not found to be within the Fungi kingdom according to data in UniProt, despite the fact that all domains in the study come from FG proteome (Figure 5-9 A). These are PF02524, PF04508, PF00746, PF00435, and PF07649 domains. This inconsistency is most likely to be due to differences in the version of PFAM and/or domain identification algorithm used to populate the UniProt database. All but one of these domains occur only once in FG proteome. The exception was PF04508, which was present in two FG proteins, namely FGSG_02793 and FGSG_16288. In total, these domains are present in seven hetero-bigrams but each of these bigrams appears only once in the FG bigram set.

Additional analysis of the fungal species-specific DUFs repertoire indicates that both non-pathogenic and plant pathogenic fungi share the same repertoire of the DUFs within their proteomes (Table 5-8).

Figure 5-10 illustrates the distribution of 35 DUFs specific to fungal species including *F. graminearum*. DUF3779 (PF12621) is the most abundant DUF that is specific to fungi. This domain is typically 100 aa in length and is likely to be involved in phosphate metabolism protein. Frequently, this DUF was found in association with DUF221 (PF02714), which was denoted earlier in this chapter (Table 5-6). DUF3779 is also very well represented within plant pathogenic fungal species accounting for 98% of the total number of plant pathogenic fungi included in this analysis (Table 5-8). However, this domain is also quite well represented within non-pathogenic fungi, where 91% of non-pathogenic fungi possess this DUF. DUF3779 is present in two FG proteins: FGSG_17175 and FGSG_05653, both with the same number, type, and order of

domains (Table 5-9). One of these domains, namely PF13967 has an assigned function as RSN1_TM, late exocytosis, associated with Golgi Transport.

A) Pfam domains



B) DUFs

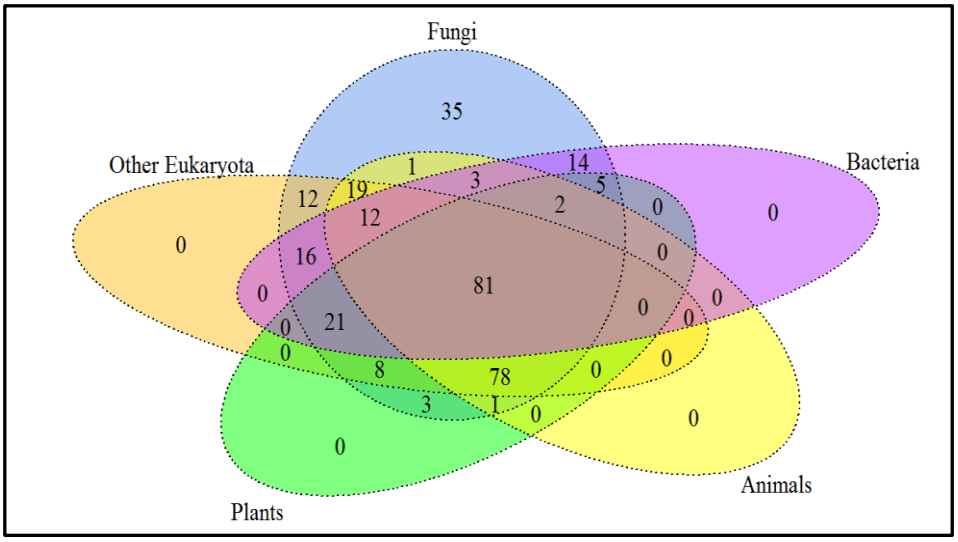


Figure 5-9 Taxonomic diversity of pfam and DUF domains occurring in *F. graminearum* proteome.
A) Taxonomic diversity of pfam domains within four taxonomic and other Eukaryotes groups. B) Taxonomic diversity of DUF domains within four taxonomic other Eukaryotes groups.

Table 5-8 DUFs specific to fungal species and their occurrence among fungi with different lifestyles.

No	DUF Id	Total number of Fungi	Plant pathogens fungi (55)			Symbiont of plants roots and endophyte (5)			Pathogens of fungi (4)			Animal pathogens fungi (61)			Non-pathogenic fungi (85)			All Pathogens (125)		
			Abundance	% of total plant pathogens	% of total fungi	Abundance	% of total symbiont of plants roots and endophyte	% of total fungi	Abundance	% of total fungi pathogens	% of total fungi	Abundance	% of total animal pathogens	% of total fungi	Abundance	% of total non-pathogens	% of total fungi	Abundance	% of total all pathogens	% of total fungi
1	DUF2456	61	23	41.82	37.70	1	20.00	1.64	3	75.00	4.92	6	9.84	9.84	28	32.94	45.90	33	26.40	54.10
2	DUF1965	72	29	52.73	40.28	2	40.00	2.78	0	0.00	0.00	14	22.95	19.44	27	31.76	37.50	45	36.00	62.50
3	DUF3716	76	29	52.73	38.16	2	40.00	2.63	2	50.00	2.63	19	31.15	25.00	24	28.24	31.58	52	41.60	68.42
4	DUF3129	83	44	80.00	53.01	4	80.00	4.82	0	0.00	0.00	16	**26.23**	19.28	19	22.35	22.89	64	26.40	77.11
5	DUF2434	84	26	47.27	30.95	2	40.00	2.38	3	75.00	3.57	26	42.62	30.95	27	31.76	32.14	57	45.60	67.86
6	DUF3176	84	34	61.82	40.48	2	40.00	2.38	3	75.00	3.57	18	29.51	21.43	27	31.76	32.14	57	45.60	67.86
7	DUF3517	96	35	63.64	36.46	2	40.00	2.08	3	75.00	3.13	30	49.18	31.25	26	30.59	27.08	70	26.40	72.92
8	DUF4045	100	36	65.45	36.00	2	40.00	2.00	3	75.00	3.00	30	49.18	30.00	29	34.12	29.00	71	56.80	71.00
9	DUF4048	101	39	70.91	38.61	2	40.00	1.98	3	75.00	2.97	29	47.54	28.71	28	32.94	27.72	73	58.40	72.28
10	DUF3636	102	38	69.09	37.25	1	20.00	0.98	3	75.00	2.94	30	49.18	29.41	30	35.29	29.41	72	26.40	70.59
11	DUF3807	103	39	70.91	37.86	1	20.00	0.97	3	75.00	2.91	27	44.26	26.21	33	38.82	32.04	70	56.00	67.96
12	DUF3984	104	39	70.91	37.50	2	40.00	1.92	3	75.00	2.88	31	50.82	29.81	29	34.12	27.88	75	60.00	72.12
13	DUF2014	108	41	74.55	37.96	2	40.00	1.85	3	75.00	2.78	27	44.26	25.00	35	41.18	32.41	73	26.40	67.59
14	DUF2457	110	41	74.55	37.27	1	20.00	0.91	3	75.00	2.73	32	52.46	29.09	33	38.82	30.00	77	61.60	70.00
15	DUF3292	111	47	85.45	42.34	2	40.00	1.80	3	75.00	2.70	23	**37.70**	20.72	36	42.35	32.43	75	60.00	67.57
16	DUF1774	119	34	61.82	28.57	2	40.00	1.68	3	75.00	2.52	38	62.30	31.93	42	49.41	35.29	77	26.40	64.71
17	DUF3328	119	43	78.18	36.13	2	40.00	1.68	3	75.00	2.52	31	50.82	26.05	40	47.06	33.61	79	63.20	66.39
18	DUF3433	119	44	80.00	36.97	3	60.00	2.52	3	75.00	2.52	31	**50.82**	26.05	38	44.71	31.93	81	64.80	68.07
19	DUF4452	120	36	65.45	30.00	1	20.00	0.83	3	75.00	2.50	35	57.38	29.17	45	52.94	37.50	75	26.40	62.50
20	DUF1770	121	43	78.18	35.54	2	40.00	1.65	3	75.00	2.48	32	52.46	26.45	41	48.24	33.88	80	64.00	66.12
21	DUF3425	133	48	87.27	36.09	2	40.00	1.50	3	75.00	2.26	36	**59.02**	27.07	44	51.76	33.08	89	71.20	66.92
22	DUF4484	139	41	74.55	29.50	2	40.00	1.44	4	100.00	2.88	39	63.93	28.06	53	62.35	38.13	86	26.40	61.87
23	DUF2011	143	38	69.09	26.57	1	20.00	0.70	4	100.00	2.80	42	68.85	29.37	58	68.24	40.56	85	68.00	59.44
24	DUF3812	147	40	72.73	27.21	2	40.00	1.36	4	100.00	2.72	40	65.57	27.21	61	71.76	41.50	86	68.80	58.50
25	DUF2406	148	41	74.55	27.70	2	40.00	1.35	4	100.00	2.70	38	62.30	25.68	63	74.12	42.57	85	26.40	57.43
26	DUF1691	148	44	80.00	29.73	4	80.00	2.70	4	100.00	2.70	41	**67.21**	27.70	55	64.71	37.16	93	74.40	62.84
27	DUF4448	150	39	70.91	26.00	4	80.00	2.67	3	75.00	2.00	39	63.93	26.00	65	76.47	43.33	85	68.00	56.67
28	DUF3115	151	42	76.36	27.81	2	40.00	1.32	4	100.00	2.65	44	72.13	29.14	59	69.41	39.07	92	26.40	60.93
29	DUF2417	155	47	85.45	30.32	2	40.00	1.29	4	100.00	2.58	40	**65.57**	25.81	62	72.94	40.00	93	74.40	60.00
30	DUF4451	159	48	87.27	30.19	4	80.00	2.52	3	75.00	1.89	46	**75.41**	28.93	58	68.24	36.48	101	80.80	63.52
31	DUF1687	160	45	81.82	28.13	5	100.00	3.13	4	100.00	2.50	39	**63.93**	24.38	67	78.82	41.88	93	26.40	58.13
32	DUF3844	162	51	92.73	31.48	4	80.00	2.47	3	75.00	1.85	48	**78.69**	29.63	56	65.88	34.57	106	84.80	65.43
33	DUF3835	165	47	85.45	28.48	5	100.00	3.03	4	100.00	2.42	43	70.49	26.06	66	77.65	40.00	99	79.20	60.00
34	DUF3602	181	53	96.36	29.28	4	80.00	2.21	4	100.00	2.21	45	73.77	24.86	75	88.24	41.44	106	26.40	58.56
35	DUF3779	188	54	98.18	28.72	3	60.00	1.60	4	100.00	2.13	50	81.97	26.60	77	90.59	40.96	111	88.80	59.04

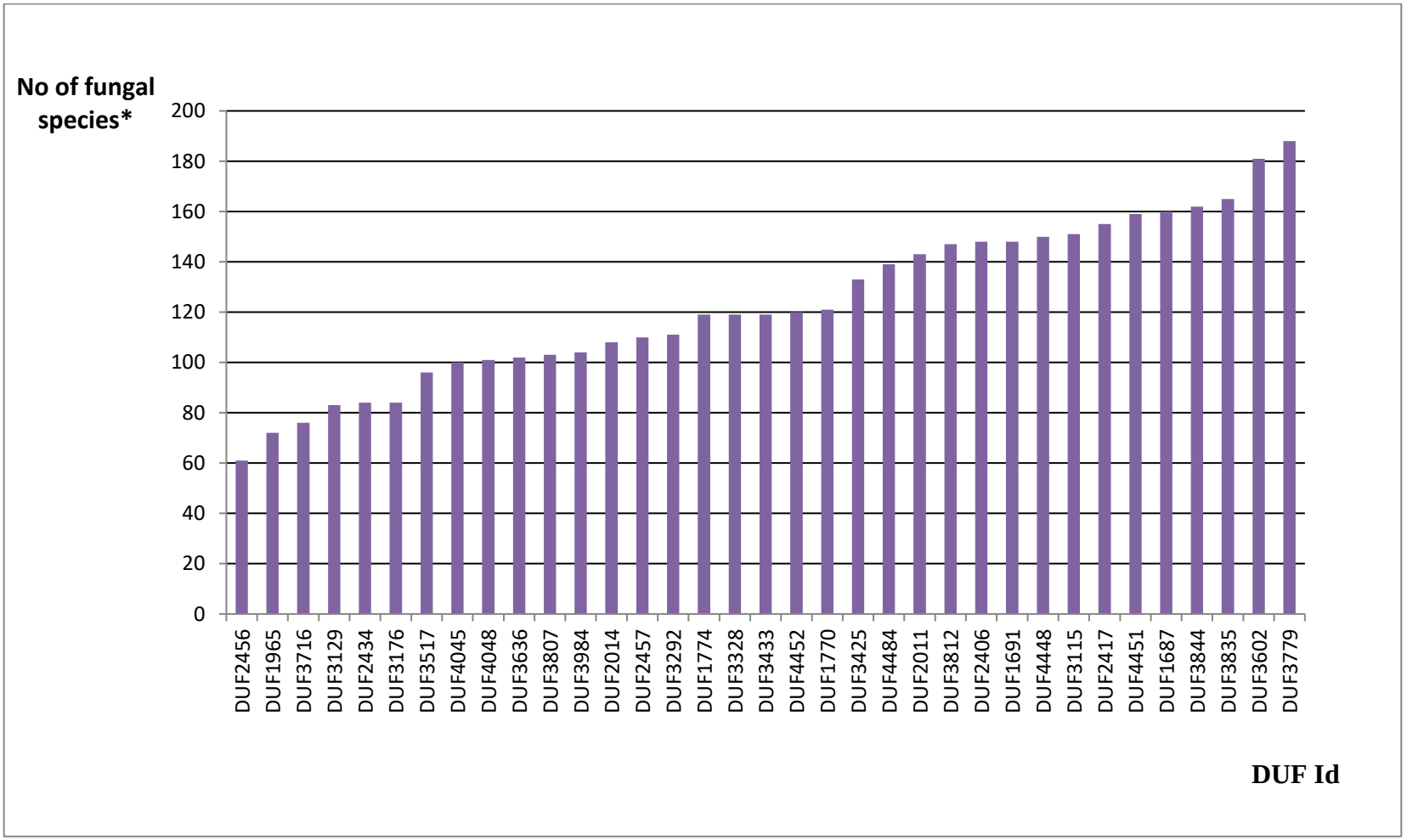


Figure 5-10 Distribution of DUFs specific to fungi.
*Number of fungal species having a given DUF

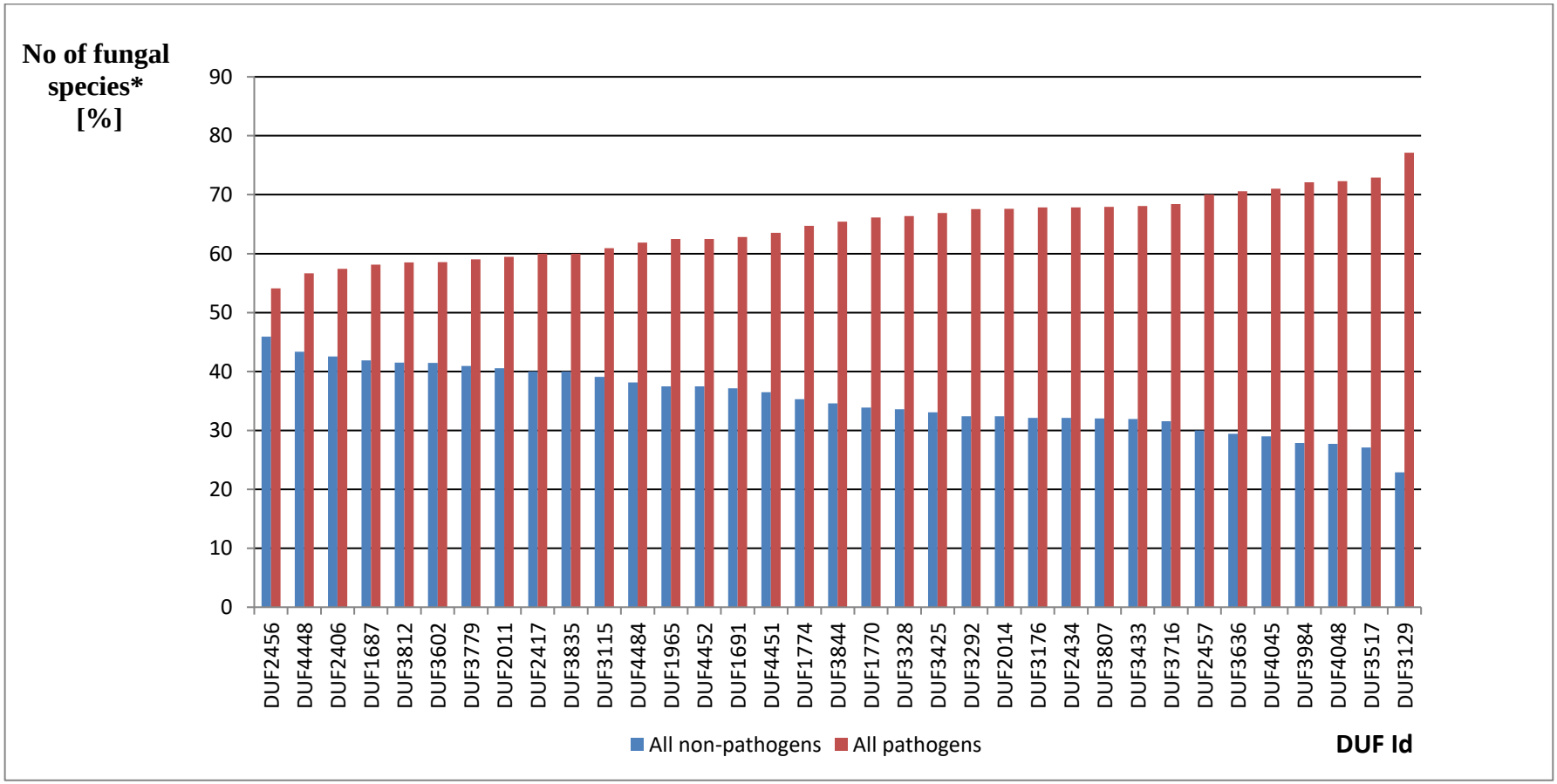


Figure 5-11 Distribution of DUFs specific to fungal species across noninfecting and infecting fungi.
A comparison is made between infecting and not infecting fungi.
*Number of fungal species having a given DUF Id.

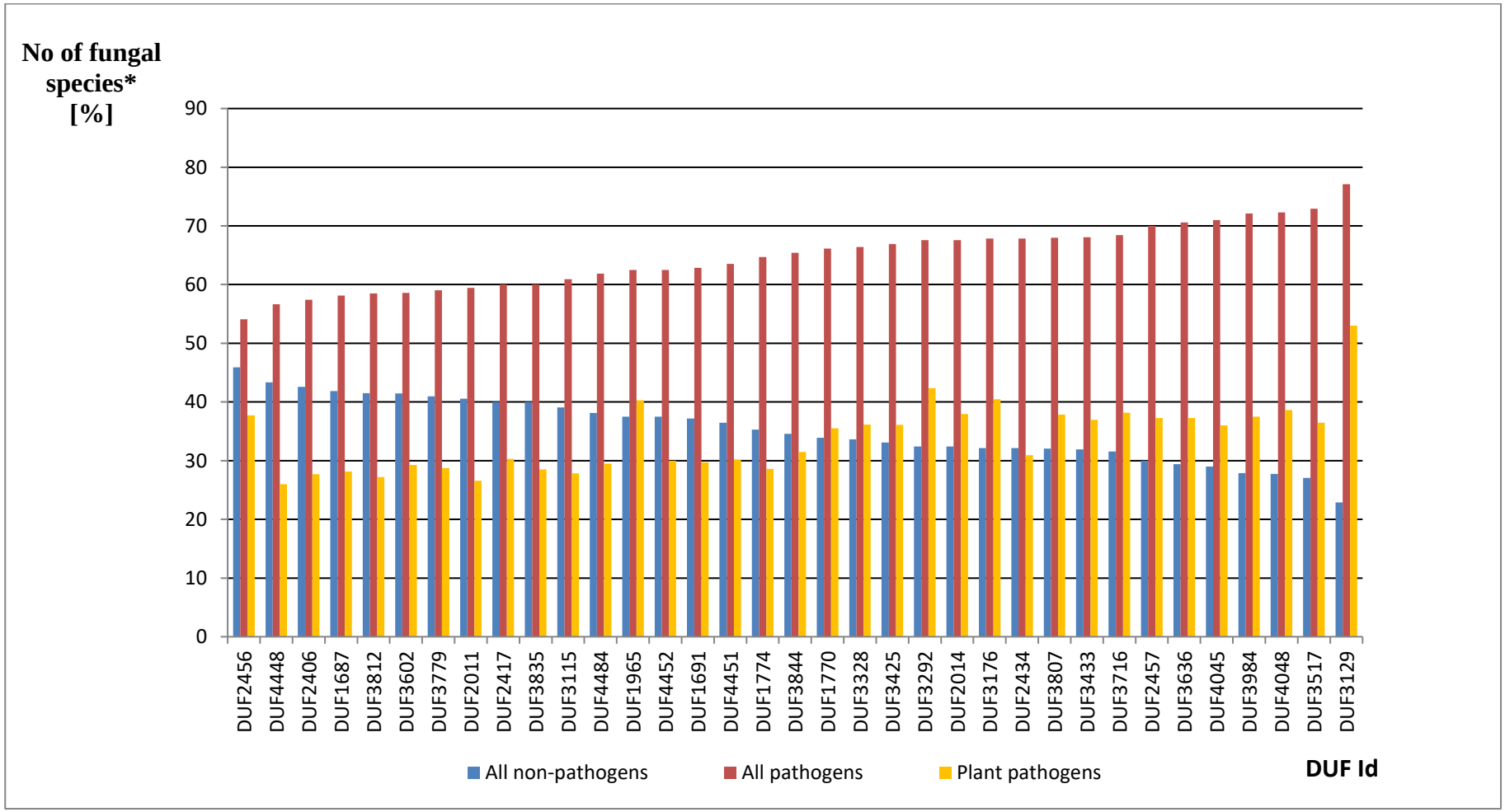


Figure 5-12 Distribution of DUFs specific to fungal species across noninfecting, infecting fungi, and fungal plant pathogens.
A comparison is made between infecting and non-infecting fungi, as well as plant fungal pathogens.
*Number of fungal species having a given DUF Id.

Table 5-9 DUFs specific to fungal species and highly enriched within plant pathogenic fungi.

DUF Id	Total no of Plant Pathogenic Fungi [%]	FG protein(s)	chromosome number	recombination frequency	Veryfied Phenotype*	predicted candidate**	bigram(s)	Protein function	One domain protein	Protein lengh [aa]	WoLF PSORT subcellular localization prediction
DUF1687	82	FGSG_08564	II	low	no	no	no	CHP	yes	139	mito: 21, cyto: 4
DUF2417	85	FGSG_15834	I	low	no	no	no	HP	yes	475	plas: 27
DUF3292	85	FGSG_16538	III	low	no	no	no	HP	yes	663	plas: 21, cyto: 3, pero: 2
		FGSG_10498	I	Middle	no	no	no	CHP	yes	654	plas: 22, E.R.: 2, cyto: 1
DUF3835	85	FGSG_00682	I	low	no	no	PF13758|DUF3835	CHP	no	611	nucl: 16.5, cyto_nucl: 13.333, mito_nucl: 9.666, cyto: 8
DUF3425	87	FGSG_04939	III	Low	no	no	no	CHP	Yes	364	mito: 13, nucl: 11, cyto: 2
		FGSG_01428	I	low	no	no	no	CHP	yes	616	nucl: 23, cyto: 3
		FGSG_13201	IV	high	no	no	no	CHP	yes	238	nucl: 16, mito: 3, cyto: 3, plas: 3
		FGSG_08284	II	middle	no	no	no	CHP	yes	291	nucl: 9.5, cyto_nucl: 8.5, cyto: 6.5, mito: 4, cysk: 4
		FGSG_11023	III	middle	no	no	no	CHP	yes	281	nucl: 14.5, cyto_nucl: 10, mito: 5, cyto: 4.5
		FGSG_09571	IV	low	no	no	no	CHP	yes	380	nucl: 19.5, cyto_nucl: 12.5, cyto: 4.5
		FGSG_01687	I	middle	no	no	no	CHP	yes	306	nucl: 18.5, cyto_nucl: 11, pero: 3, cyto: 2.5
		FGSG_12345	II	middle	yes (UP)	no	PF00170|DUF3425	CHP	no	497	nucl: 26.5, cyto_nucl: 14
		FGSG_06824	IV	low	no	no	no	CHP	yes	467	nucl: 18, mito: 4, cyto: 3
		FGSG_16025	I	low	no	no	no	HP	yes	510	nucl: 21, cyto: 5
		FGSG_03230	II	middle	no	no	no	CHP	yes	469	nucl: 23, cyto: 2
		FGSG_06968	IV	low	no	no	no	CHP	yes	356	nucl: 21, pero: 2, mito: 1, cyto: 1
DUF4451	87	FGSG_16699	IV	low	no	no	no	CHP	yes	608	nucl: 24.5, cyto_nucl: 13.5
DUF3844	93	FGSG_11820	I	low	no	no	no	CHP	yes	371	extr: 12, E.R.: 5, mito: 4, cyto: 3, golg: 2
DUF3602	96	FGSG_01224	I	low	no	no	no	CHP	yes	152	nucl: 15, cyto_nucl: 15, cyto: 9
		FGSG_10386	I	low	no	no	no	CHP	yes	138	mito: 13.5, cyto_mito: 9.833, nucl: 6.5, cyto_nucl: 6.333, cyto: 5
		FGSG_11936	I	low	no	no	no	CHP	yes	132	nucl: 17, cyto_nucl: 15, cyto: 9
		FGSG_06726	IV	low	no	no	no	CHP	yes	164	nucl: 14, cyto_nucl: 12.5, cyto: 9, mito: 1
DUF3779	98	FGSG_17175	IV	low	no	no	PF13967|DUF4463 DUF4463|DUF221 DUF221|DUF3779	(related to RSN1 – Overexpression rescues sro7/sop1 in NaCl	no	1038	plas: 18, E.R.: 7
		FGSG_05653	III	low	no	no	PF13967|DUF4463 DUF4463|DUF221 DUF221|DUF3779	related to *A.thaliana* hyp1 protein	no	897	plas: 26

Where HP – hypothetical protein, CHP – conserved hypothetical protein, UP – unaffected pathogenicity; *phenotype verified based on the information in PHI-base version 3.4; **candidate genes for pathogenicity predicted in (Lysenko et al., 2013). Where cellular compartments are cyto – cytosol, cysk – cytoskeleton, E.R.- endoplasmic reticulum, extr - extracellular location, golg - Golgi apparatus, pero - peroxisome, plas - plasma membrane, mito – mitochondria, nucl – nucleus, vacu – vacuole.

DUF3779 is not part of the main component of the domain-domain network (section 5.4.4 of this chapter). Both proteins FGSG_17175 and FGSG_05653 (Table 5-9) are present within a low recombination region of the genome and neither of them has so far been associated with virulence or pathogenicity (PHI-base version 3.4 and 3.6) (Urban et al., 2015b) of FG nor predicted to be candidate virulence genes in the previous study (Lysenko et al., 2013).

The second most enriched fungal-specific DUF present in FG is DUF3602 (PF12223). The domain is typically between 78 and 89 amino acids in length and it is present in 181 out of the total of 210 different fungal species. This domain is very well represented within plant pathogenic fungal species accounting for 96% of the total number of plant fungal pathogens included in this analysis (Table 5-8). However, it is also quite well represented within non-pathogenic fungi (88%). DUF3602 is present as a single pfam domain in four different FG conserved hypothetical proteins, which are within the low recombination regions of the FG genome. These four FG proteins are neither associated with the virulence or pathogenicity (PHI-base version 3.4 and 3.6) (Urban et al., 2015b) nor candidate genes predicted in the previous study (Lysenko et al., 2013).

Parallel analysis of the fungal species-specific DUFs repertoire among fungi with different lifestyle (Table 5-8 and Figure 5-11) reveals an enrichment of all 35 DUFs within pathogenic fungi comparing to non-pathogenic fungi. Two of the most abundant DUFs were found to be specific to pathogenic fungal species, namely DUF3129 (PF11327) and DUF3517 (PF12030).

DUF3129 is present in 83 out of 210 different fungal species included in this analysis. It is also well represented in plant pathogenic fungal species accounting for 80% of the total number of plant-infecting fungi (Table 5-8). On the other hand, the occurrence of this DUF within non–pathogenic fungi is only 22%. DUF3129 is very well represented within symbionts of plants roots and endophytes accounting for 80% of this group (Table 5-8).

DUF3129 is also the only domain present in four FG proteins: FGSG_04647 (gene located on chromosome II in a high recombination region, length 295 aa), FGSG_09353 (gene located on chromosome IV in a low recombination region, length 404 aa), FGSG_00006 (gene located on chromosome I in a high recombination region, length 296 aa), and FGSG_08021 (gene located

on chromosome II in a low recombination region, length 289 aa). First three of them are predicted to be Egh16 proteins. The function, however, relates to the entire protein and not necessary the domain comprising the protein. Although Egh16 may play important role during the early infection stage in fungal pathogens (Xue et al., 2002), none of the four FG proteins is associated with the virulence or pathogenicity (PHI-base version 3.4 and 3.6) (Urban et al., 2015b) nor candidate genes predicted in the previous study (Lysenko et al., 2013).

The DUF3517 has a predicted length of 340 aa. DUF3517 is present in 96 out of the total of 210 different species in this analysis. This DUF is very well represented in 3 fungi that infect other fungi (75%). It is also well represented in plant pathogenic fungi, where 64% of plant pathogenic fungal species have this DUF within their proteome. By comparison, only 31% of non-pathogenic fungal species have this DUF in their proteome. A similar trend was observed (Table 5-8) when comparing the percentage of total number of fungi which are plant pathogens (36%) to the percentage of the total number of fungi which are not pathogenic (27%). DUF3517 is present in only one FG protein namely FGSG_17669 (related to a DEFRY protein) forming a bigram with PFAM domain PF00443 (Ubiquitin carboxyl-terminal hydrolase). FGSG_17669 is a long protein (2587 aa), codes for by a gene residing in a very low recombination region of FG chromosome III. FGSG_17669 is neither associated with the virulence or pathogenicity (PHI-base version 3.4 and 3.6) (Urban et al., 2015b) nor candidate gene predicted in the previous study (Lysenko et al., 2013).

Although DUFs are more abundant among all pathogenic fungi, further analysis of DUF enrichment in plant-infecting fungi in contrast to non-pathogenic fungal species failed to confirm the earlier observation for all DUFs (Figure 5-11). Only about 49% (17 out of 35) of the DUFs present in plant-infecting fungi verify the earlier finding that DUFs are more enriched within host infecting fungal species. However, for 9 of these 17 DUFs the enrichment in plant-infecting fungi is higher than 80% (Table 5-8). Further examination of this sub-set of DUFs (Table 5-9) revealed that most of these DUFs were the only domain so far predicted within the given FG protein. The exceptions here are DUF3835, DUF3425, and DUF3779, where these fungi-specific DUFs form bigrams with other pfam domains or with the other DUFs, as per DUF3779 previously described (see above). Although all DUFs listed in Table 5-9 are highly enriched in plant-infecting fungi, the

majority of them are associated with FG protein for which experimentally verified phenotype information is absent (PHI-base version 3.4 and 3.6 (Urban et al., 2015b)). The exception to this is DUF3425 which is present in 12 FG proteins of which FGSG_12345 is associated with unaffected pathogenicity phenotype according to PHI-base version 3.6 (Urban et al., 2015b). This protein has also been identified as an FG transcription factor by a previous study (Son et al., 2011). Furthermore, no FG proteins, in which those highly abundant in plant-infecting fungi DUFs are present, have been predicted as candidate genes by Lysenko's study (Lysenko et al., 2013).

This extended analysis of the 35 DUFs specific to fungal species demonstrates that their whole repertoire is present in nine plant pathogenic fungi, two fungi infecting other fungi, one endophyte fungus and in one saprotrophic fungus (Table 5-10). This indicates that these 35 DUFs are present in fungal species with different pathogenic lifestyle, i.e. the ability to infect and colonise cereal as well as non-cereal hosts. Further examination of the nine plant pathogenic fungi, with regards to the tree of life for *Fusaria*, revealed a very tight cluster of five fungi with one slight outlier (*N. haematococca, also known as Fusarium solani*), one slightly distant species (*Colletotrichum gloeosporioides and Colletotrichum graminicola)* and more distant to above *Magnaporthe oryzae*. However, this group of plant pathogens all reside within Ascomycota and all of them are possessing a classical hemi-biotrophic lifestyle.

Table 5-10 Fungi with whole repertoire of DUFs specific to fungal species.

No	Plant pathogens possessing all 35 DUFs	Fungi infecting fungi with all 35 DUFs	Symbionts of plants roots and endophyte	Not infecting Fungi with all 35 DUFs
1	*Colletotrichum gloeosporioides*	*Hypocrea atroviridis*	*Pestalotiopsis fici*	*Podospora anserina*
2	*Colletotrichum graminicola*	*Hypocrea virens*		
3	*Fusarium oxysporum*			
4	*Fusarium pseudograminearum*			
5	*Gibberella fujikuroi*			
6	*Gibberella moniliformis*			
7	*Gibberella zeae*			
8	*Magnaporthe oryzae*			
9	*Nectria haematococca*			

It is not understood why the saprotrophic species *Podospora anserina* has the entire repertoire of 35 DUFs. This outcome might suggest that this fungus has not yet been experimentally tested on immune compromised hosts for pathogenicity. This might also suggest that the species could

have been taxonomically misclassified within the Sordariomycetes or that this fungal species has acquired novel genes from pathogenic fungal species via horizontal gene transfer.

As expected, no DUF or set of DUFs was found to be specific to *Fusarium* and FG only. Thus, there is no evidence that a DUF which is unique to other kingdoms has been horizontally transferred only into FG.

5.4.3.1 Statistical evaluation of DUFs association with fungi lifestyle

Inspecting the data presented in Table 5-8, one general question immediately arises: Do the 35 fungi-specific DUFs either occur independently across the fungi with a different lifestyle or is there a correlation between specific DUF and a particular fungal lifestyle? To answer this question the chi-square test of association was chosen with the null hypothesis, H0, assuming that 35 fungi-specific DUFs occur independently amongst the fungi with a different lifestyle.

The chi-square test was not conducted for the original categories (as per Table 5-8) such as PP (plant pathogens), SP (symbionts and endophyte), FP (pathogen of fungi), AP (animal pathogens) and NP (non-pathogens). This was due to the condition of the test, where there should not be more than 20% of expected values less than five in the test frequency table. In fact, all expected values for SP and FP categories are less than five (Appendix C Table C-1). The issue was initially resolved by merging SP and FP values together into SPFP group, then combining them with other pathogenic categories into other pathogens (OP), PP+SPFP and all pathogens (AllPath) groups and finally exclude SP and FP groups from the chi-square test.

In total six chi-square tests were performed for several combinations of the lifestyle categories. These are test 1 (testing PP, SPFP, AP and NP categories), test 2 (testing PP, OP and NP categories), test 3 (examining PP+SPFP, AP and NP sets), test 4 (testing PP, AP and NP sets), test 5 (examining PP and NP sets), and test 6 (testing AllPath and NP sets). Generated contingency tables for all tests are displayed in Table 5-13. The results for each chi-square test are presented in Table 5-11 and tests frequency tables, including original categories table, are displayed in Appendix C as: Table C-1, Table C-2, Table C-3, Table C-4, Table C-5, Table C-6, and Table C-7.

Table 5-11 Chi-square tests result summary.

Chi-square test	χ^2	Critical value	d.f.	p-value	α
test 1	93.823	126.57	102	0.7059	0.05
test 2	84.753	88.25	68	0.0824	0.05
test 3	85.851	88.25	68	0.0708	0.05
test 4	88.585	88.25	68	0.0476	0.05
test 5	61.591	48.60	34	0.0026	0.05
test 6	49.153	48.60	34	0.0448	0.05

χ2 – chi-square test statistic value, α – uncertainty level, d.f. – degree of freedom.

Table 5-12 Discrepancy comparison.

DUF Id	(O-E)2/E						
	Test 4			Test 5		Test 6	
	PP	AP	NP	PP	NP	AllPath	NP
DUF2406	(-) 1.24	(-) 0.07	(+) 1.68	(-) 1.49	(+) 1.36	(-) 1.00	(+) 1.77
DUF3129	(+) 10.49	(-) 1.68	(-) 3.90	(+) 6.49	(-) 5.91	(+) 2.23	(-) 3.96
DUF3292	(+) 3.08	(-) 1.49	(-) 0.39	(+) 1.39	(-) 1.27	(+) 0.22	(-) 0.39
DUF3517	(+) 0.45	(+) 0.81	(-) 2.01	(+) 1.20	(-) 1.10	(+) 1.19	(-) 2.12
DUF4048	(+) 1.10	(+) 0.17	(-) 1.85	(+) 1.56	(-) 1.42	(+) 1.08	(-) 1.92
DUF4448	(-) 2.08	(-) 0.03	(+) 2.29	(-) 2.26	(+) 2.06	(-) 1.26	(+) 2.24

Where PP- plant pathogenic fungi, AP – fungi infecting animals, AllPath– includes fungi lifestyle groups: plant pathogenic fungi (PP), symbionts of plant roots and endophyte (SP), fungi infecting other fungi (FP) and fungi infecting animals (AP); NP – non-pathogenic fungi. Signs: (+) or (-) indicate positive or negative associations with the given fungal lifestyle categories.

The null hypothesis stating that fungal-specific DUFs would appear independently within different lifestyle was confirmed by three tests: test 1, test 2 and test 3 (Table 5-11). However, when testing the homogeneity of 35 DUFs amongst PP, AP, NP and AllPath categories the null hypothesis was rejected with the highest significant result obtained while testing association of each of 35 fungal-specific DUFs with two categories of fungal lifestyles: PP and NP (test 5 (Figure 5-11), p-value = 0.0026). In addition, the null hypothesis was rejected in two further tests reported in Table 5-11, namely test 4 (p-value = 0.0476) and test 6 (p-value = 0.0448). Inspection of the frequency tables for the test 4, test 5, and test 6 (Appendix C: Table C-5, Table C-6 and Table C-7 respectively) reveals that the greatest discrepancy from expected values was observed for DUF3129 and DUF4448. This is very well notable while studying the frequency table for test 4 (Appendix C, Table C-5), as well as Table 5-12, where discrepancy for DUF3129 is the highest (10.49) towards PP and discrepancy for DUF4448 is the highest (2.29) for NP.

Table 5-13 Chi-square test contingency tables.

Table	Original table						test 1					test 2				test 3				test 4				test 5			test 6		
DUF Id	PP	SP	FP	AP	NP	total	PP	SPFP	AP	NP	total	PP	OP	NP	total	PP+SPFP	AP	NP	total	PP	AP	NP	total	PP	NP	total	AllPath	NP	total
DUF2456	23	1	3	6	28	61	23	4	6	28	61	23	10	28	61	27	6	28	61	23	6	28	57	23	28	51	33	28	61
DUF1965	29	2	0	14	27	72	29	2	14	27	72	29	16	27	72	31	14	27	72	29	14	27	70	29	27	56	45	27	72
DUF3716	29	2	2	19	24	76	29	4	19	24	76	29	23	24	76	33	19	24	76	29	19	24	72	29	24	53	52	24	76
DUF3129	44	4	0	16	19	83	44	4	16	19	83	44	20	19	83	48	16	19	83	44	16	19	79	44	19	63	64	19	83
DUF2434	26	2	3	26	27	84	26	5	26	27	84	26	31	27	84	31	26	27	84	26	26	27	79	26	27	53	57	27	84
DUF3176	34	2	3	18	27	84	34	5	18	27	84	34	23	27	84	39	18	27	84	34	18	27	79	34	27	61	57	27	84
DUF3517	35	2	3	30	26	96	35	5	30	26	96	35	35	26	96	40	30	26	96	35	30	26	91	35	26	61	70	26	96
DUF4045	36	2	3	30	29	100	36	5	30	29	100	36	35	29	100	41	30	29	100	36	30	29	95	36	29	65	71	29	100
DUF4048	39	2	3	29	28	101	39	5	29	28	101	39	34	28	101	44	29	28	101	39	29	28	96	39	28	67	73	28	101
DUF3636	38	1	3	30	30	102	38	4	30	30	102	38	34	30	102	42	30	30	102	38	30	30	98	38	30	68	72	30	102
DUF3807	39	1	3	27	33	103	39	4	27	33	103	39	31	33	103	43	27	33	103	39	27	33	99	39	33	72	70	33	103
DUF3984	39	2	3	31	29	104	39	5	31	29	104	39	36	29	104	44	31	29	104	39	31	29	99	39	29	68	75	29	104
DUF2014	41	2	3	27	35	108	41	5	27	35	108	41	32	35	108	46	27	35	108	41	27	35	103	41	35	76	73	35	108
DUF2457	41	1	3	32	33	110	41	4	32	33	110	41	36	33	110	45	32	33	110	41	32	33	106	41	33	74	77	33	110
DUF3292	47	2	3	23	36	111	47	5	23	36	111	47	28	36	111	52	23	36	111	47	23	36	106	47	36	83	75	36	111
DUF1774	34	2	3	38	42	119	34	5	38	42	119	34	43	42	119	39	38	42	119	34	38	42	114	34	42	76	77	42	119
DUF3328	43	2	3	31	40	119	43	5	31	40	119	43	36	40	119	48	31	40	119	43	31	40	114	43	40	83	79	40	119
DUF3433	44	3	3	31	38	119	44	6	31	38	119	44	37	38	119	50	31	38	119	44	31	38	113	44	38	82	81	38	119
DUF4452	36	1	3	35	45	120	36	4	35	45	120	36	39	45	120	40	35	45	120	36	35	45	116	36	45	81	75	45	120
DUF1770	43	2	3	32	41	121	43	5	32	41	121	43	37	41	121	48	32	41	121	43	32	41	116	43	41	84	80	41	121
DUF3425	48	2	3	36	44	133	48	5	36	44	133	48	41	44	133	53	36	44	133	48	36	44	128	48	44	92	89	44	133
DUF4484	41	2	4	39	53	139	41	6	39	53	139	41	45	53	139	47	39	53	139	41	39	53	133	41	53	94	86	53	139
DUF2011	38	1	4	42	58	143	38	5	42	58	143	38	47	58	143	43	42	58	143	38	42	58	138	38	58	96	85	58	143
DUF3812	40	2	4	40	61	147	40	6	40	61	147	40	46	61	147	46	40	61	147	40	40	61	141	40	61	101	86	61	147
DUF2406	41	2	4	38	63	148	41	6	38	63	148	41	44	63	148	47	38	63	148	41	38	63	142	41	63	104	85	63	148
DUF1691	44	4	4	41	55	148	44	8	41	55	148	44	49	55	148	52	41	55	148	44	41	55	140	44	55	99	93	55	148
DUF4448	39	4	3	39	65	150	39	7	39	65	150	39	46	65	150	46	39	65	150	39	39	65	143	39	65	104	85	65	150
DUF3115	42	2	4	44	59	151	42	6	44	59	151	42	50	59	151	48	44	59	151	42	44	59	145	42	59	101	92	59	151
DUF2417	47	2	4	40	62	155	47	6	40	62	155	47	46	62	155	53	40	62	155	47	40	62	149	47	62	109	93	62	155
DUF4451	48	4	3	46	58	159	48	7	46	58	159	48	53	58	159	55	46	58	159	48	46	58	152	48	58	106	101	58	159
DUF1687	45	5	4	39	67	160	45	9	39	67	160	45	48	67	160	54	39	67	160	45	39	67	151	45	67	112	93	67	160
DUF3844	51	4	3	48	56	162	51	7	48	56	162	51	55	56	162	58	48	56	162	51	48	56	155	51	56	107	106	56	162
DUF3835	47	5	4	43	66	165	47	9	43	66	165	47	52	66	165	56	43	66	165	47	43	66	156	47	66	113	99	66	165
DUF3602	53	4	4	45	75	181	53	8	45	75	181	53	53	75	181	61	45	75	181	53	45	75	173	53	75	128	106	75	181
DUF3779	54	3	4	50	77	188	54	7	50	77	188	54	57	77	188	61	50	77	188	54	50	77	181	54	77	131	111	77	188
Total:	1418	84	109	1155	1556	4322	1418	193	1155	1556	4322	1418	1348	1556	4322	1611	1155	1556	4322	1418	1155	1556	4129	1418	1556	2974	2766	1556	4322

PP - plant pathogenic fungi, SP –symbionts of plant roots and endophyte, FP – fungi infecting other fungi, AP – fungi infecting animals, NP – non - pathogenic fungi, SPFP – combined SP and FP sets, OP – other pathogens (combined SP, FP and AP sets), O- observed frequencies, E – expected frequencies

In addition, while inspecting the frequency tables for test 5 (Appendix C, Table C-6) and test 6 (Appendix C, Table C-7) a similar pattern for both DUFs was observed. This suggests that DUF3129 is positively correlated with plant fungal pathogens (PP), whereas DUF4448 is positively correlated with non-pathogenic fungi (NP). Moreover, there are further three DUFs for which positive correlation with plant pathogenic fungi (PP) has been identified (Table 5-12). These are DUF3292 (the strongest correlation with PP for the test 4), DUF3517 (the strongest correlation with PP for the test 5) and DUF4048 (the strongest correlation with PP for the test 5). In addition, DUF2406 has been also identified as only associated with non-pathogenic fungi (Table 5-12) with slightly lower values of discrepancy comparing to DUF4448 in favour of NP throughout test 4, test 5, and test 6.

5.4.4 *Fusarium graminearum* domain-association network

A total of 314 unique DUFs have been identified within FG proteome. Of these, 101 of DUFs have been formally associated with the hetero-bigrams formation. In addition, those 101 DUFs are part of the pfam domain-association network where domains are nodes and edges represent bigrams of hetero (different) domains (see method section 5.3.4). The graphical visualisation of the network is presented in Figure 5-14 and its main metrics are listed in Table 5-14.

There are 386 connected components (CCs) in the network and 267 of them consist of only two nodes (see Appendix C Table C-8). The DUFs distribution among the whole network was found to be uneven (Appendix C Table C-8). The DUFs appear in 48 CCs of the network which accounts for 12% of the total number of CCs in the network. There are 11 CCs with DUF only nodes (Figure 5-13). These are components with two nodes only. The largest CC consists of 40 unique DUFs (Table 5-14), which accounts for nearly 40% of unique DUFs spread across the whole network. The most highly connected DUF in the largest connected component is DUF1929, which is present in five FG proteins: FGSG_09093, FGSG_03569, FGSG_00251, FGSG_11097 and FGSG_11032. These proteins were previously identified as either related to galactose oxidase precursor (FGSG_09093, FGSG_11097), or probable galactose oxidase (FGSG_03569 and FGSG_00251), or GAOA galactose oxidase Precursor (FGSG_11032).

As expected, pfam domains with which DUF1929 forms bigrams are in the most of the cases Kelch motif (PF13418, PF01344, PF07646) and in one case glyoxal oxidase N-terminus domain (PF07250). None of the five FG proteins was associated either with the virulence or pathogenicity (PHI-base version 3.6) (Urban et al., 2015b) or previously predicted to be associated with a virulence phenotype (Lysenko et al., 2013).

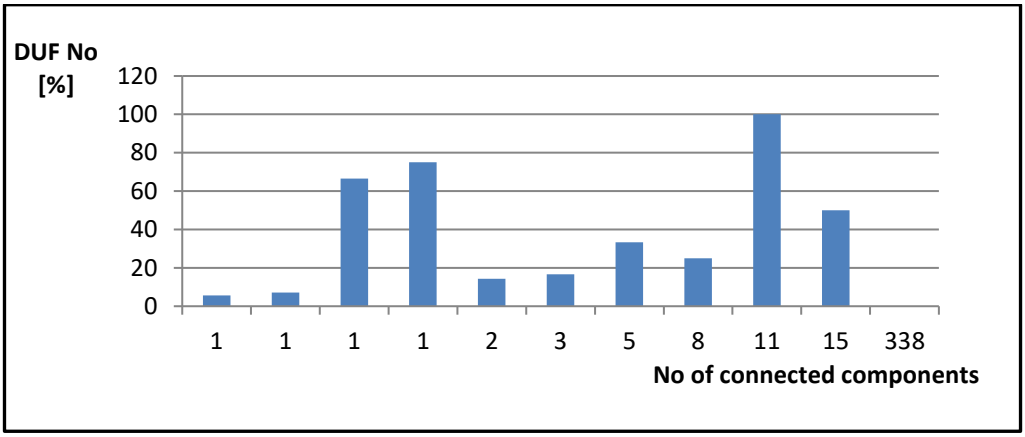
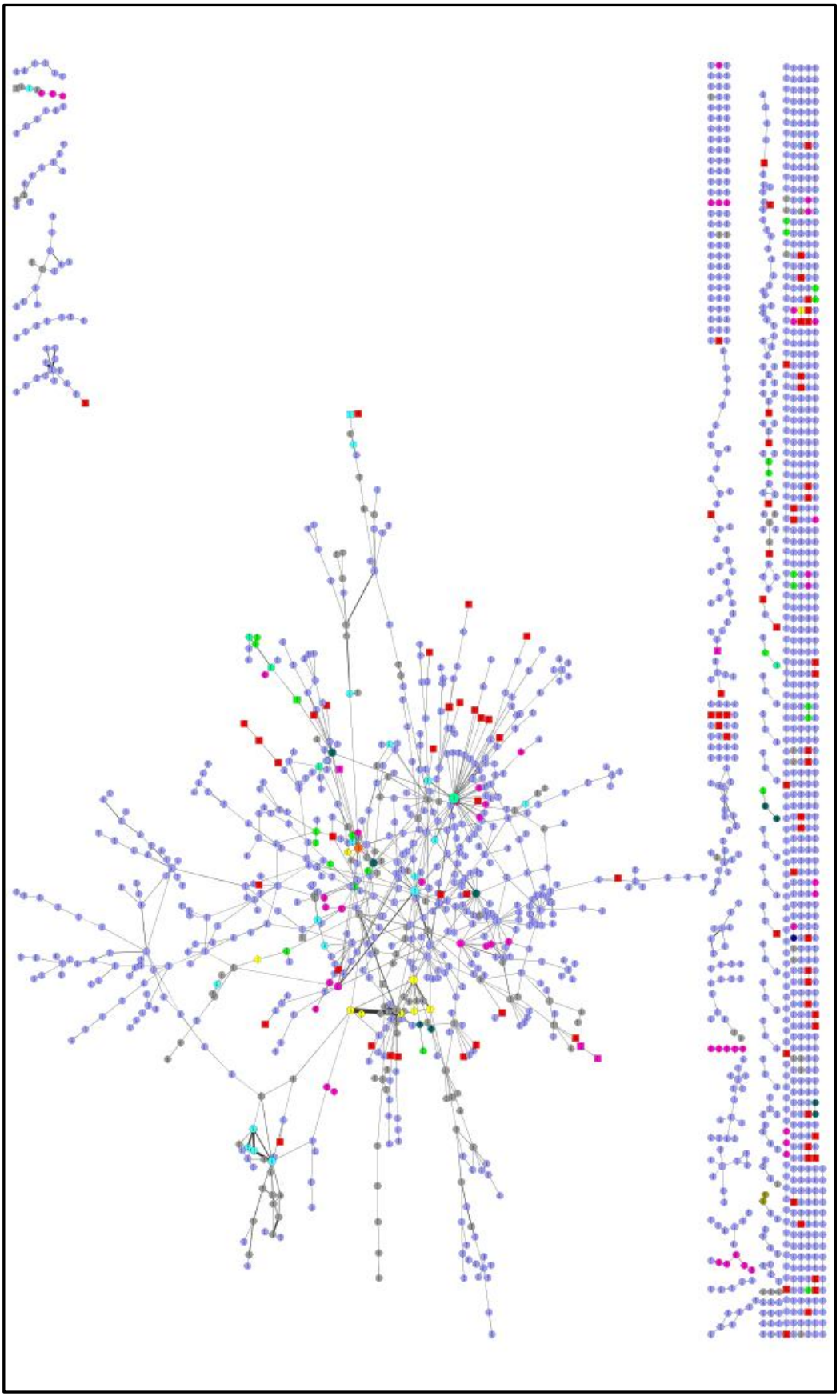


Figure 5-13 DUFs content among connected components of the domain-association network.

Table 5-14 Main properties of the domain-association network compared with properties of the largest connected component of that network.

Main metrics	The whole network	The largest connected component
Total nodes number	1747	705
Number of DUF nodes	101	40
Edges number	1523	849
Number of connected components	383	1
Network diameter	20	20
Network centralisation	0.018	0.045
Network density	0.001	0.003
Average number of neighbours	1.744	2.409
Most highly connected node	PF00400 (node degree = 34)	PF00400 (node degree = 34)
Average clustering coefficient	0.045	0.076
Network heterogeneity	1.063	1.1
Number of self-loops	0	0
Multi-edge node pairs	0	0
Modularity	0.9434	0.8784



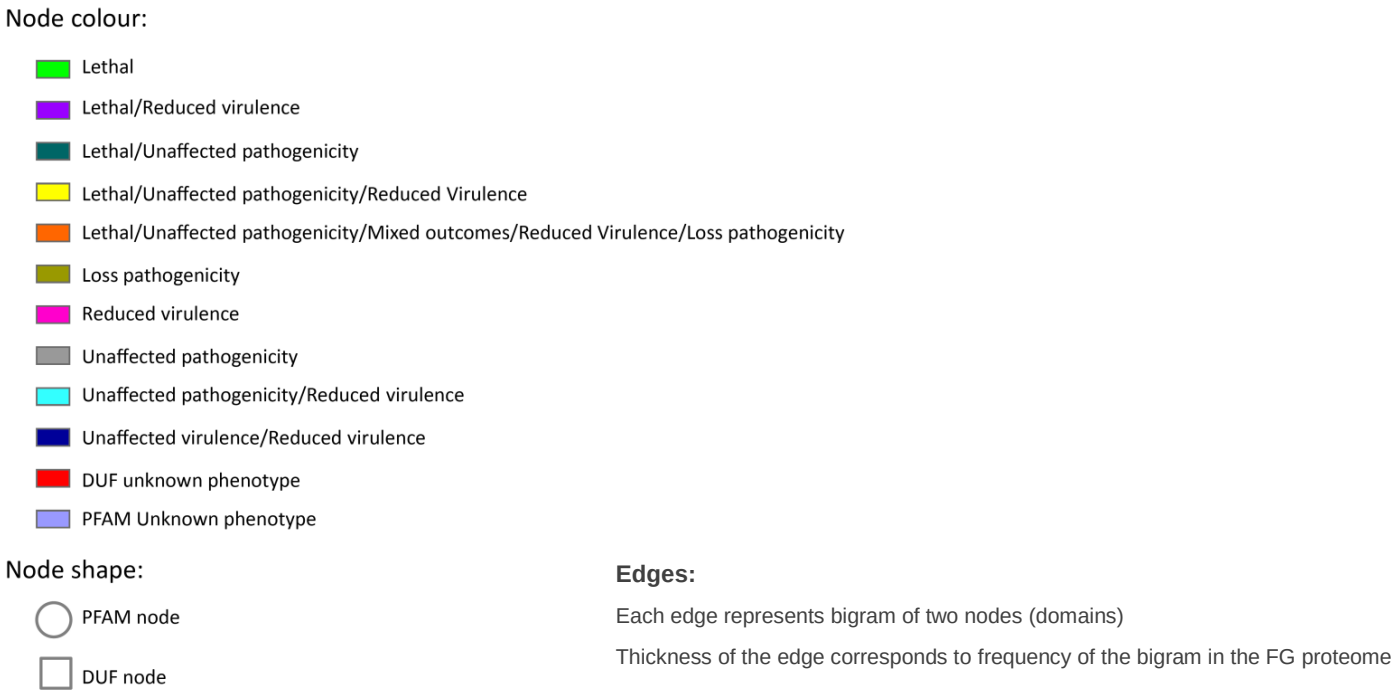


Figure 5-14 Graphical representation of the domain-association network.
DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either it is a DUF (square) node or pfam node (circle).

5.4.5 Nodes topological properties statistics

Based on the information from PHI-base (version 3.4, released on 12 Feb 2013) there are 938 PHI-base entries associated with the FG proteins for which the role on the pathogenicity outcome was experimentally tested. Out of those entries, 912 unique FG proteins were assigned the mutant phenotype including as follows: reduced virulence, increased virulence, loss pathogenicity, lethal, mixed outcomes, as well as unaffected pathogenicity and unaffected virulence.

In this study, based on the phenotype information associated with FG protein, domains building these proteins inherited the phenotype from the protein. Thus, some domains are associated with several different phenotypic outcomes. These domains appear in several proteins linked with different phenotypes. Whereas some domains show only one phenotype. Figure 5-14 presents domains with different phenotypes linked to a FG protein. Those domains are highlighted with different colours in the network (see Figure 5-14 legend for details). In this section the main properties of a node such as node degree, clustering coefficient and degree centrality were separately calculated for each node. Nodes were grouped into five categories: all nodes except DUFs, DUF nodes, lethal nodes (except DUFs), reduced virulence and loss of pathogenicity nodes (except DUFs), as well as unaffected pathogenicity nodes (except DUFs). Here we are interested if DUF nodes have either similar topological property to the nodes associated with virulence (reduced virulence and loss of pathogenicity nodes), lethal nodes or to nodes associated with proteins not influencing the pathogenicity (unaffected pathogenicity nodes).

5.4.5.1 Distributions comparison

Despite being fundamental to data visualisation, histograms can often be a poor method for comparing the distribution (see Appendix C, Figures C-3 to C-6). On the other hand, a Kernel Density Plot (KDP) can be a more effective way to show and compare the distribution of variables. KDP visualises the distribution of data over a continuous interval or time. In other words, KDP is a non-parametric way to estimate the probability density function of a random variable. Appendix C Figures C-7 to C-12 depict KDPs of examined properties of nodes for a studied nodes group. KDPs were generated with the aid of R software version 3.03.

Examining KDPs depicted in Appendix C (Figures C-7 to C-12), it can be noted that the distribution of DUF nodes for each tested property of a node follows the one for all nodes. In addition, slightly similar distribution was also observed for nodes associated with virulence proteins (reduced virulence and loss pathogenicity nodes (Appendix C: Figure C-7 D, Figure C-9 D and Figure C-11 D). In contrasts, DUF-node distribution was found not to follow the distribution of unaffected pathogenicity nodes (Appendix C: Figure C-7 E, Figure C-9 E and Figure C-11 E). This is very well demonstrated while testing a node clustering coefficient (Appendix C Figure C-9 E.). This finding might indicate that DUFs are associated with the pathogenic process.

5.4.5.2 Statistical tests

Non-parametric statistical tests were performed to compare the distribution of DUF nodes to all nodes (excluding DUFs) and other nodes groups such as reduced virulence, loss pathogenicity, reduced virulence and loss pathogenicity, lethal and unaffected pathogenicity nodes. The first test, Wilcoxon-Mann-Whitney rank sum test (equivalent to Wilcoxon Rank Sum Test with continuity correction in R package) was used to compare averages of two independent tested groups. In three tested node parameters, it was observed that significant differences ($p < 0.05$) existed between most tested node groups and DUF nodes (Table 5-15 A, Table 5-16 A and Table 5-17 A).

However, no significant difference was observed between the lethal and DUF node distributions in all three tested node parameters: the node degree (p-value = 0.1061), the clustering coefficient (p-value = 0.9323) and the node degree centrality (p-value = 0.1061) (Table 5-15 A, Table 5-16 A and Table 5-17 A).

In addition, no significant difference was observed between the distributions of the loss pathogenicity and DUF nodes for all tested node parameters: node degree (p-value = 0.2711), node clustering coefficient (p-value = 0.8026), and node degree centrality (p-value = 0.2711).

Table 5-15 Node degree distributions comparison.

A) Wilcoxon-Mann-Whitney rank sum test (equivalent to Wilcoxon Rank Sum Test with continuity correction in R package)

NODE DEGREE										
WILCOXON RANK SUM TEST with continuity correction, two sided										
NODES		SAMPLE SIZE		MEDIAN		95% CONFIDENCE LEVEL				
1	2	1	2	1	2	W	p-value	95% confidence interval		difference in location
All Nodes -DUF	DUF	1646	101	1	1	98897	1.3180E-04	1.2480E-05	8.0641E-05	1.1304E-05
LethalOnly-DUF	DUF	23	101	1	1	1337.5	1.0610E-01	-4.6361E-05	5.6053E-05	1.5841E-05
RedVirOnly - DUF	DUF	51	101	2	1	3566.5	1.4260E-06	4.5376E-06	9.9997E-01	1.5983E-05
LossPathOnly - DUF	DUF	2	101	1.5	1	132.5	2.7110E-01	-5.8281E-06	1.0000E+00	3.9253E-05
redVirOnly+LossPathOnly -DUF	DUF	53	101	2	1	3699	1.3290E-06	7.1187E-05	9.9996E-01	2.5301E-05
UnaffPathOnly - DUF	DUF	132	101		1	9530.5	1.4160E-10	6.1223E-05	9.9998E-01	9.9999E-01
NODES		SAMPLE SIZE		MEDIAN		99% CONFIDENCE LEVEL				
1	2	1	2	1	2	W	p-value	99% confidence interval		difference in location
All Nodes -DUF	DUF	1646	101	1	1	98897	1.3180E-04	3.2696E-05	1.9005E-05	1.1304E-05
Lethal Only	DUF	23	101	1	1	1337.5	1.0610E-01	-5.2474E-08	5.7332E-05	1.5841E-05
RedVirOnly - DUF	DUF	51	101	2	1	3566.5	1.4260E-06	4.7235E-05	9.9998E-01	1.5983E-05
LossPathOnly - DUF	DUF	2	101	1.5	1	132.5	2.7110E-01	-6.6785E-06	1.0000E+00	3.9253E-05
redVirOnly+LossPathOnly -DUF	DUF	53	101	2	1	3699	1.3290E-06	2.8350E-05	9.9993E-01	2.5301E-05
UnaffPathOnly - DUF	DUF	132	101	2	1	9530.5	1.4160E-10	2.6077E-05	1.0001E+00	9.9999E-01

B) Kolmogorov-Smirnov test

NODE DEGREE							
Kolmogorov-Smirnov Test							
NODES		SAMPLE SIZE		MEDIAN			
1	2	1	2	1	2	D	p-value
All Nodes -DUF	DUF	1646	101	1	1	0.1717	7.3060E-03
Lethal Only	DUF	23	101	1	1	0.1541	7.6520E-01
RedVirOnly - DUF	DUF	51	101	2	1	0.3512	4.6840E-04
LossPathOnly - DUF	DUF	2	101	1.5	1	0.3218	9.8720E-01
redVirOnly+LossPathOnly -DUF	DUF	53	101	2	1	0.3501	3.9880E-04
UnaffPathOnly - DUF	DUF	132	101	2	1	0.3748	2.0840E-07

1 – first data set;
2 – second dataset
All Nodes - DUF – the set contains all nodes except DUF nodes
Lethal Only – DUF – the set contains only lethal nodes (no mixture of lethal and other phenotypes nodes) without DUFs being associated with lethal phenotype
RedVirOnly – DUF – the set contains only reduced virulence nodes (no mixture of reduced virulence and other phenotypes nodes) without DUFs being associated with reduced virulence phenotype
LossPathOnly - DUF - the set contains only loss pathogenicity nodes (no mixture of loss pathogenicity node and other phenotypes nodes) without DUFs being associated with loss pathogenicity phenotype
redVirOnly+LossPathOnly – DUF - the set contains only reduced virulence and loss pathogenicity nodes (no mixture of those nodes with other phenotypes nodes) without DUFs being associated with reduced virulence and loss pathogenicity phenotype
UnaffPathOnly – DUF - the set contains unaffected pathogenicity nodes (no mixture of those nodes with other phenotypes nodes) without DUFs being associated with unaffected pathogenicity phenotype

Table 5-16 Node degree clustering coefficient comparison.

A) Wilcoxon-Mann-Whitney rank sum test (equivalent to Wilcoxon Rank Sum Test with continuity correction in R package)

Node Clustering Coefficient										
WILCOXON RANK SUM TEST with continuity correction, two sided										
NODES		SAMPLE SIZE		MEDIAN		95% CONFIDENCE LEVEL				
1	2	1	2	1	2	W	p-value	95% confidence interval		difference in location
All Nodes -DUF	DUF	1646	101	0	0	87237.5	8.8160E-02	-1.9009E-07	2.5130E-05	8.2077E-05
LethalOnly-DUF	DUF	23	101	0	0	1166.5	9.3230E-01	-2.2926E-05	3.1830E-05	1.6167E-05
RedVirOnly - DUF	DUF	51	101	0	0	2865.5	1.5690E-02	-5.6634E-05	6.5516E-05	3.0358E-05
LossPathOnly - DUF	DUF	2	101	0	0	97	8.0260E-01	0.0000E+00	0.0000E+00	0.0000E+00
redVirOnly+LossPathOnly -DUF	DUF	53	101	0	0	2962.5	1.9490E-02	-7.2529E-05	5.4184E-05	3.1603E-05
UnaffPathOnly - DUF	DUF	132	101	0	0	8028.5	2.6790E-05	5.3316E-05	2.2695E-06	1.7761E-05
NODES		SAMPLE SIZE		MEDIAN		99% CONFIDENCE LEVEL				
1	2	1	2	1	2	W	p-value	99% confidence interval		difference in location
All Nodes -DUF	DUF	1646	101	0	0	87237.5	8.8160E-02	-4.8563E-05	3.3390E-05	8.2077E-05
Lethal Only	DUF	23	101	0	0	1166.5	9.3230E-01	-5.8208E-05	1.1171E-05	1.6167E-05
RedVirOnly - DUF	DUF	51	101	0	0	2865.5	1.5690E-02	-8.6998E-05	2.2771E-05	3.0358E-05
LossPathOnly - DUF	DUF	2	101	0	0	97	8.0260E-01	0.0000E+00	0.0000E+00	0.0000E+00
redVirOnly+LossPathOnly -DUF	DUF	53	101	0	0	2962.5	1.9490E-02	-5.1326E-05	1.5180E-05	3.1603E-05
UnaffPathOnly - DUF	DUF	132	101	0	0	8028.5	2.6790E-05	1.8744E-05	2.5321E-05	1.7761E-05

B) Kolmogorov-Smirnov test

Node Clustering Coefficient							
Kolmogorov-Smirnov Test							
NODES		SAMPLE SIZE		MEDIAN			
1	2	1	2	1	2	D	p-value
All Nodes -DUF	DUF	1646	101	0	0	0.0509	9.6600E-01
Lethal Only	DUF	23	101	0	0	0.0138	1.0000E+00
RedVirOnly - DUF	DUF	51	101	0	0	0.1173	7.4000E-01
LossPathOnly -DUF	DUF	2	101	0	0	0.0396	1.0000E+00
redVirOnly+LossPathOnly -DUF	DUF	53	101	0	0	0.1113	7.8200E-01
UnaffPathOnly - DUF	DUF	132	101	0	0	0.2104	1.2620E-02

1 – first data set;
2 – second dataset
All Nodes - DUF – the set contains all nodes except DUF nodes
Lethal Only – DUF – the set contains only lethal nodes (no mixture of lethal and other phenotypes nodes) without DUFs being associated with lethal phenotype
RedVirOnly – DUF – the set contains only reduced virulence nodes (no mixture of reduced virulence and other phenotypes nodes) without DUFs being associated with reduced virulence phenotype
LossPathOnly - DUF - the set contains only loss pathogenicity nodes (no mixture of loss pathogenicity node and other phenotypes nodes) without DUFs being associated with loss pathogenicity phenotype
redVirOnly+LossPathOnly – DUF - the set contains only reduced virulence and loss pathogenicity nodes (no mixture of those nodes with other phenotypes nodes) without DUFs being associated with reduced virulence and loss pathogenicity phenotype
UnaffPathOnly – DUF - the set contains unaffected pathogenicity nodes (no mixture of those nodes with other phenotypes nodes) without DUFs being associated with unaffected pathogenicity phenotype

Table 5-17 Node degree centralities comparison.

A) Wilcoxon-Mann-Whitney rank sum test (equivalent to Wilcoxon Rank Sum Test with continuity correction in R package)

Node Degree Centrality										
WILCOXON RANK SUM TEST with continuity correction, two sided										
NODES		SAMPLE SIZE		MEDIAN		95% CONFIDENCE LEVEL				
1	2	1	2	1	2	W	p-value	95% confidence interval		difference in location
All Nodes -DUF	DUF	1646	101	5.7274E-04	5.7274E-04	98896.5	1.3180E-04	7.9911E-05	5.1298E-05	3.3162E-05
LethalOnly-DUF	DUF	23	101	5.7274E-04	5.7274E-04	1337.5	1.0610E-01	-1.6469E-05	5.9038E-05	2.9615E-05
RedVirOnly - DUF	DUF	51	101	1.1455E-03	5.7274E-04	3566.5	1.4260E-06	6.1993E-05	5.5551E-04	5.5230E-05
LossPathOnly - DUF	DUF	2	101	8.5911E-04	5.7274E-04	132.5	2.7110E-01	-1.3498E-06	5.7274E-04	2.5464E-05
redVirOnly+LossPathOnly -DUF	DUF	53	101	1.1455E-03	5.7274E-04	3699	1.3290E-06	5.4638E-05	5.4328E-04	4.2821E-05
UnaffPathOnly - DUF	DUF	132	101	1.1455E-03	5.7274E-04	9530.5	1.4160E-10	6.9379E-06	5.6858E-04	5.1057E-04
NODES		SAMPLE SIZE		MEDIAN		99% CONFIDENCE LEVEL				
1	2	1	2	1	2	W	p-value	99% confidence interval		difference in location
All Nodes -DUF	DUF	1646	101	5.7274E-04	5.7274E-04	98896.5	1.3180E-04	3.5312E-05	1.0210E-05	3.3162E-05
Lethal Only	DUF	23	101	5.7274E-04	5.7274E-04	1337.5	1.0610E-01	-4.0328E-05	5.0641E-05	2.9615E-05
RedVirOnly - DUF	DUF	51	101	1.1455E-03	5.7274E-04	3566.5	1.4260E-06	8.0920E-05	5.4922E-04	5.5230E-05
LossPathOnly - DUF	DUF	2	101	8.5911E-04	5.7274E-04	132.5	2.7110E-01	-9.7442E-06	5.7274E-04	2.5464E-05
redVirOnly+LossPathOnly -DUF	DUF	53	101	1.1455E-03	5.7274E-04	3699	1.3290E-06	7.8933E-05	5.3026E-04	4.2821E-05
UnaffPathOnly - DUF	DUF	132	101	1.1455E-03	5.7274E-04	9530.5	1.4160E-10	4.8098E-08	6.3456E-04	5.1057E-04

B) Kolmogorov-Smirnov test

Node Degree Centrality							
Kolmogorov-Smirnov Test							
NODES		SAMPLE SIZE		MEDIAN			
1	2	1	2	1	2	D	p-value
All Nodes -DUF	DUF	1646	101	5.7274E-04	5.7274E-04	0.1717	7.3060E-03
Lethal Only	DUF	23	101	5.7274E-04	5.7274E-04	0.1541	7.6520E-01
RedVirOnly - DUF	DUF	51	101	1.1455E-03	5.7274E-04	0.3512	4.6840E-04
LossPathOnly -DUF	DUF	2	101	8.5911E-04	5.7274E-04	0.3218	9.8720E-01
redVirOnly+LossPathOnly -DUF	DUF	53	101	1.1455E-03	5.7274E-04	0.3501	3.9880E-04
UnaffPathOnly - DUF	DUF	132	101	1.1455E-03	5.7274E-04	0.3748	2.0840E-07

1 – first data set;
2 – second dataset
All Nodes - DUF – the set contains all nodes except DUF nodes
Lethal Only – DUF – the set contains only lethal nodes (no mixture of lethal and other phenotypes nodes) without DUFs being associated with lethal phenotype
RedVirOnly – DUF – the set contains only reduced virulence nodes (no mixture of reduced virulence and other phenotypes nodes) without DUFs being associated with reduced virulence phenotype
LossPathOnly - DUF - the set contains only loss pathogenicity nodes (no mixture of loss pathogenicity node and other phenotypes nodes) without DUFs being associated with loss pathogenicity phenotype
redVirOnly+LossPathOnly – DUF - the set contains only reduced virulence and loss pathogenicity nodes (no mixture of those nodes with other phenotypes nodes) without DUFs being associated with reduced virulence and loss pathogenicity phenotype
UnaffPathOnly – DUF - the set contains unaffected pathogenicity nodes (no mixture of those nodes with other phenotypes nodes) without DUFs being associated with unaffected pathogenicity phenotype

However, the sample size for loss pathogenicity nodes is very low (it is equal to 2) and this might explain the lack of significance in that case. There was also found a weak significant difference between the distributions of DUF nodes and other node sets such as all nodes (except DUF), reduced virulence nodes (except DUF), reduced virulence and loss of pathogenicity nodes (except DUF), with p-values equal to 0.08816, 0.01569, 0.01949 respectively when testing for the node clustering coefficient (Table 5-16 A).

Another non-parametric statistical test, the Kolmogorov-Smirnov (KS) test was carried out to confirm the initial finding. The KS test is a much stronger test than Wilcoxon-Mann-Whitney rank sum test. As previously found, no significant difference was observed between the distributions of lethal and DUF nodes in all three tested node parameters: the node degree (D = 0.1541, p-value = 0.7652), the clustering coefficient (D = 0.0138, p-value = 1) and the node degree centrality (D = 0.1541, p-value = 0.7652) (Table 5-15 B, Table 5-16 B and Table 5-17 B). Moreover, loss pathogenicity nodes distribution was found not to be significantly different to DUF-node distribution for three tested node parameters: the node degree (D = 0.3218, p-value = 0.9872), the node clustering coefficient (D = 0.0396, p-value = 1) and the node degree centrality (D =0.3218, p-value= 0.9872). However, the very small sample size of loss pathogenicity nodes can explain this lack of the significance difference. The KS test also confirmed that there was also no significant difference between distributions of DUF nodes and other node sets such as all nodes (except DUFs) (D = 0.0509, p-value = 0.966), reduced virulence nodes, except DUFs, (D = 01173, p-value = 0.740), reduced virulence and loss of pathogenicity nodes, except DUFs, (D = 0.1113, p-value = 0.782) when testing for the node clustering coefficient parameter (Table 5-16 B). In addition, KS test revealed a weak significant difference between distributions of unaffected pathogenicity and DUF nodes (D= 0.2104, p-value = 0.01262) when testing for the node clustering coefficient parameter.

5.4.6 Articulation points calculation

This type of analysis focusses on the largest connected component of the domain-association network (Figure 5-15). A cut vertex (articulation point) is any vertex (node) whose removal increases the number of connected components.

As indicated earlier, the largest connected component of the network consists of 705 interacting domains. This accounts for 40% of the total number of domains within the whole network. There were 259 cut vertices (articulation points) identified in the largest connected component of the network (Table 5-18).

Although articulation points account for nearly 38% of the total number of vertices within the largest connected component, removing any of them did not lead to the substantial change in the largest connected component topology. In the majority of instances, removing one of the cut vertices from the largest connected component resulted in one large connected component with nodes number ranging from 668 (in case of removal PF00400 node) to 703 and several small connected components with nodes number ranging from 1 to 17. Of the most interest are articulation points whose removal leads to substantial increase in the number of CCs. Figure 5-16 illustrates the example when the node PF00400 was removed from the largest connected component of the domain-association network. This results in splitting up the largest connected component of the domain-association network into one large connected component consisting of 668 nodes, 5 small connected components (one with eight nodes, one with three nodes, and three with two nodes) and 19 unconnected nodes including four DUF nodes (DUF3337, DUF1900, DUF3639 and DUF1899).

Another example of a cut vertex removal from the largest connected component of the domain-association network is presented in Figure 5-17. Here, we observe the largest connected component splitting up into two connected components: one with 687 nodes and the second with 17 nodes. As seen from Table 5-18, there were eight DUF cut vertexes identified within the largest CC of domain-association network. After a thorough examination of each DUF cut vertex, removal any of them did not affect the topology of the CC and in the majority led to the disconnection of the only one node from the main CC (five out of eight cases). In three out of five instances, the node separated from the main CC was a DUF node. Moreover, removing following DUF articulation points: DUF4414, DUF1929 and DUF3385 from the largest CC resulted in disconnection of two, three and eight nodes respectively from the largest CC. Figure 5-18 illustrates the example when removal of DUF3386 from the main CC resulted in one large CC with 696 nodes and one small CC with eight nodes.

Overall, a removal of any articulation point (cut vertex) from the largest connected component of the domain-association network did not result in drastic changes in size and topology of the largest connected component.

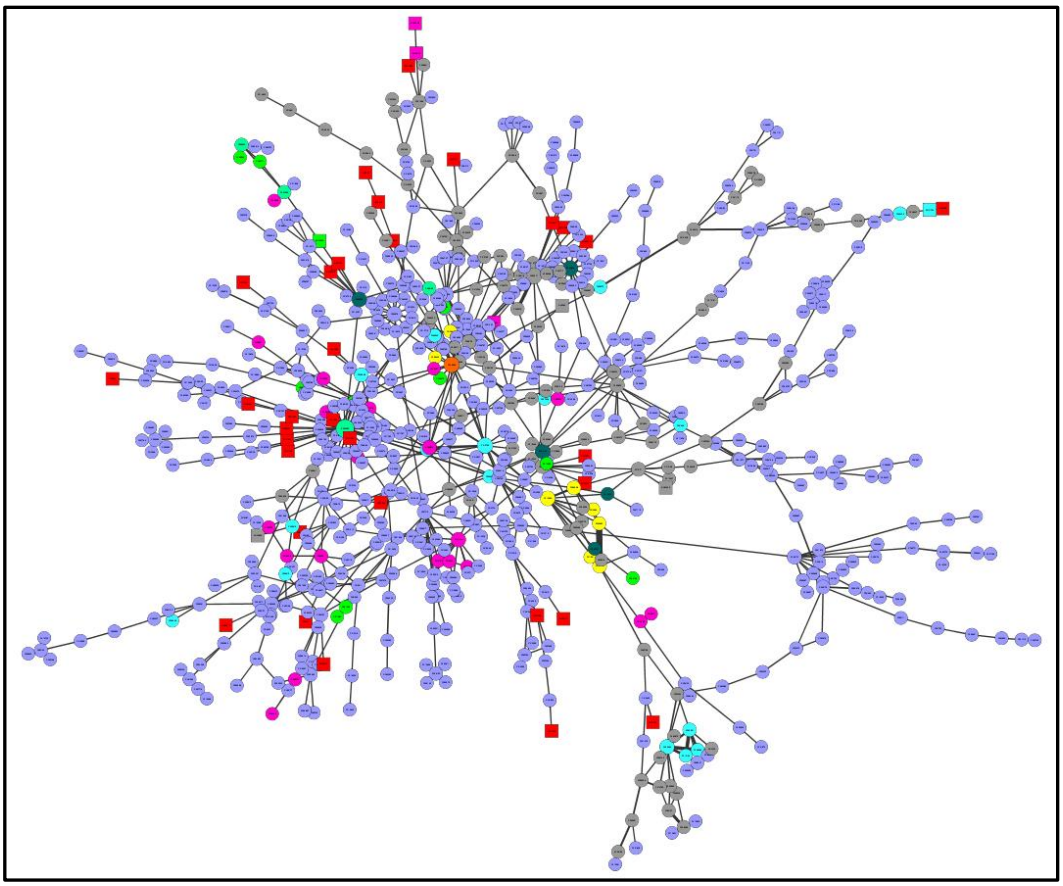


Figure 5-15 The largest connected component of the domain-association network.
DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either it is a DUF (square) node or pfam node (circle).

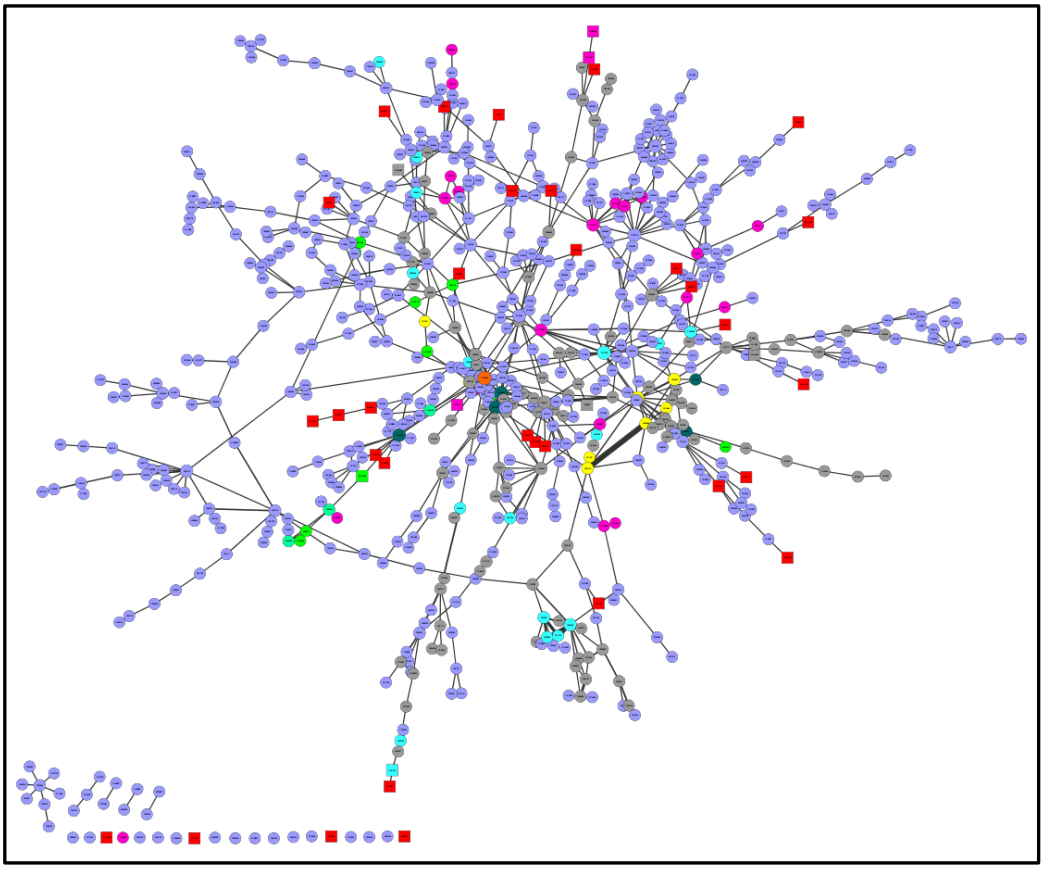


Figure 5-16 The influence on the topology of the largest connected component of the domain-association network after the removal of PF00400 node.
DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either it is a DUF (square) node or pfam node (circle).

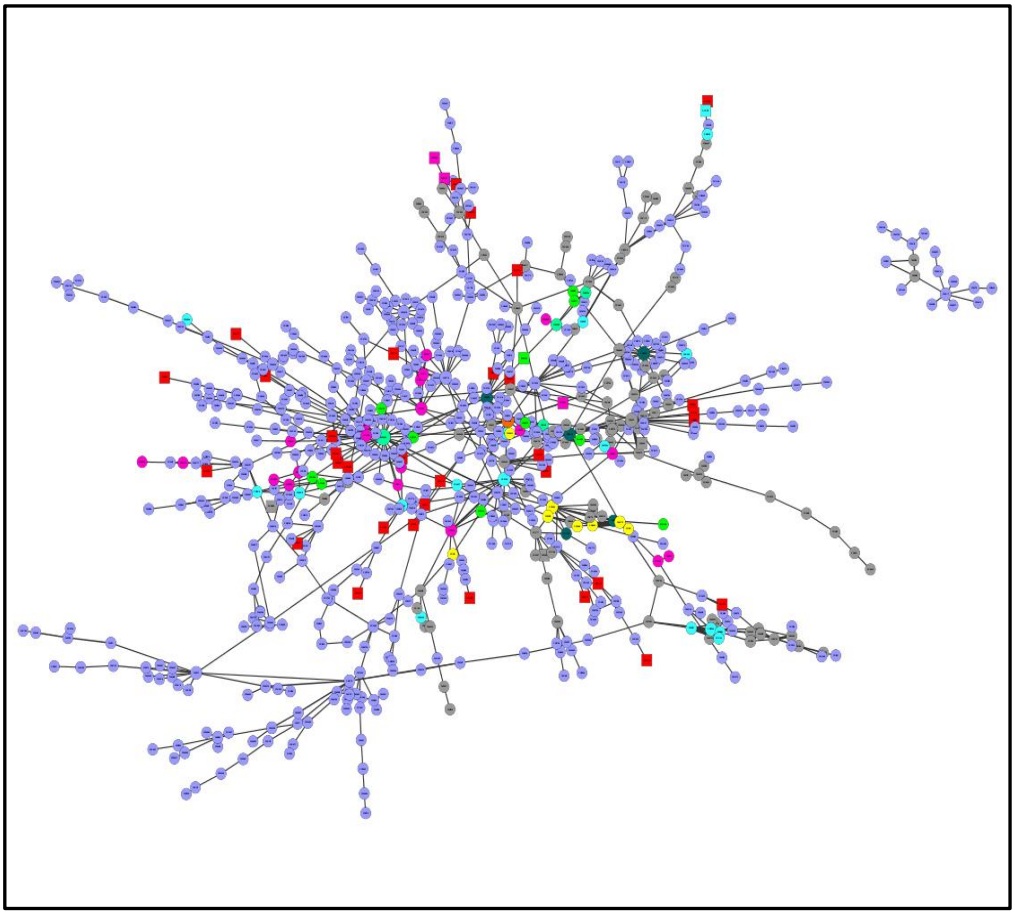


Figure 5-17 The influence on the topology of the largest connected component of the domain-association network after the removal of PF02785 node.
DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either it is a DUF (square) node or pfam node (circle).

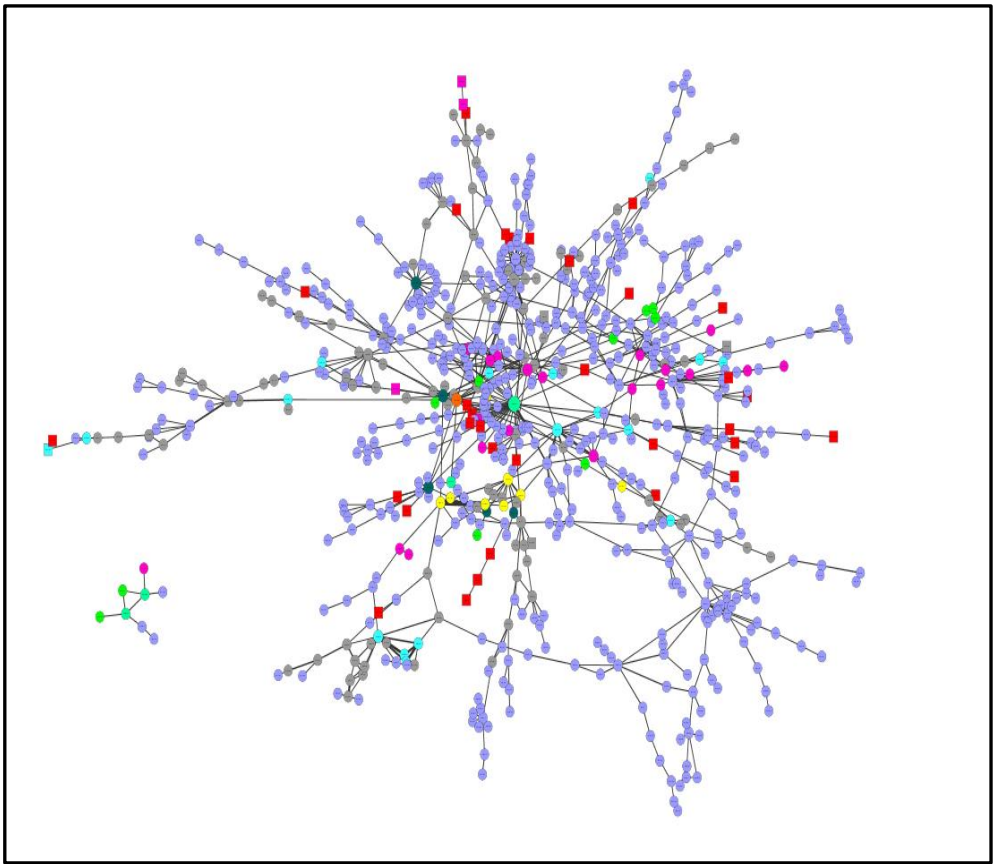


Figure 5-18 The influence on the topology of the largest connected component of the domain-association network after the removal of DUF3385 node.
DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either it is a DUF (square) node or pfam node (circle).

Table 5-18 List of cut vertices (articulation points) of the largest connected component of the domain-association network.

No	node ID	CC No	No	node ID	CC No	No	node ID	CC No	No	node ID	CC No	No	node ID	CC No
1	PF00400	25	**53**	PF13191	3	**105**	PF14479	2	**157**	PF04928	2	**209**	PF00989	2
2	PF00271	13	**54**	PF08241	3	**106**	PF14327	2	**158**	PF04926	2	**210**	PF00941	2
3	PF00004	13	**55**	PF08240	3	**107**	PF14259	2	**159**	PF04898	2	**211**	PF00916	2
4	PF00076	12	**56**	PF07500	3	**108**	PF13857	2	**160**	PF04851	2	**212**	PF00899	2
5	PF02985	9	**57**	PF04969	3	**109**	PF13855	2	**161**	PF04715	2	**213**	PF00890	2
6	PF00069	9	**58**	PF03221	3	**110**	PF13838	2	**162**	PF04679	2	**214**	PF00790	2
7	PF00173	8	**59**	PF02883	3	**111**	PF13646	2	**163**	PF04564	2	**215**	PF00788	2
8	PF02178	7	**60**	PF02809	3	**112**	PF13519	2	**164**	PF04508	2	**216**	PF00786	2
9	PF00226	7	**61**	PF02801	3	**113**	PF13452	2	**165**	PF04433	2	**217**	PF00780	2
10	PF00117	7	**62**	PF02786	3	**114**	PF13428	2	**166**	PF04408	2	**218**	PF00754	2
11	PF13920	6	**63**	PF02770	3	**115**	PF13426	2	**167**	PF04389	2	**219**	PF00698	2
12	PF13894	6	**64**	PF02142	3	**116**	PF13424	2	**168**	PF04053	2	**220**	PF00636	2
13	PF04055	6	**65**	PF01753	3	**117**	PF13415	2	**169**	PF03876	2	**221**	PF00632	2
14	PF01585	6	**66**	PF01751	3	**118**	PF13405	2	**170**	PF03810	2	**222**	PF00620	2
15	PF00787	6	**67**	PF01602	3	**119**	PF13393	2	**171**	PF03460	2	**223**	PF00618	2
16	PF00533	6	**68**	PF01424	3	**120**	PF13236	2	**172**	PF03450	2	**224**	PF00613	2
17	PF00443	6	**69**	PF01388	3	**121**	PF13181	2	**173**	PF03368	2	**225**	PF00612	2
18	PF00085	6	**70**	PF01363	3	**122**	PF12874	2	**174**	PF03142	2	**226**	PF00611	2
19	PF13418	5	**71**	PF01302	3	**123**	PF12807	2	**175**	PF03127	2	**227**	PF00610	2
20	PF13086	5	**72**	PF00970	3	**124**	PF12763	2	**176**	PF02933	2	**228**	PF00583	2
21	PF12171	5	**73**	PF00642	3	**125**	PF12464	2	**177**	PF02893	2	**229**	PF00570	2
22	PF07719	5	**74**	PF00637	3	**126**	PF12436	2	**178**	PF02877	2	**230**	PF00569	2
23	PF02518	5	**75**	PF00628	3	**127**	PF11764	2	**179**	PF02785	2	**231**	PF00498	2
24	PF00856	5	**76**	PF00616	3	**128**	PF10607	2	**180**	PF02771	2	**232**	PF00487	2
25	PF00646	5	**77**	PF00571	3	**129**	PF10585	2	**181**	PF02729	2	**233**	PF00481	2
26	PF00575	5	**78**	PF00560	3	**130**	PF10366	2	**182**	PF02666	2	**234**	PF00441	2
27	PF00557	5	**79**	PF00550	3	**131**	PF09453	2	**183**	PF02383	2	**235**	PF00415	2
28	PF00176	5	**80**	PF00454	3	**132**	PF09382	2	**184**	PF02225	2	**236**	PF00388	2
29	PF00149	5	**81**	PF00439	3	**133**	PF09359	2	**185**	PF02204	2	**237**	PF00387	2
30	PF13513	4	**82**	PF00397	3	**134**	PF09110	2	**186**	PF02185	2	**238**	PF00385	2
31	PF13414	4	**83**	PF00364	3	**135**	PF09070	2	**187**	PF02138	2	**239**	PF00307	2
32	PF12796	4	**84**	PF00289	3	**136**	PF09011	2	**188**	PF02136	2	**240**	PF00270	2
33	PF07992	4	**85**	PF00258	3	**137**	PF08799	2	**189**	PF02020	2	**241**	PF00218	2
34	PF07653	4	**86**	PF00249	3	**138**	PF08766	2	**190**	PF01822	2	**242**	PF00204	2
35	PF05739	4	**87**	PF00240	3	**139**	PF08711	2	**191**	PF01799	2	**243**	PF00172	2
36	PF05729	4	**88**	PF00175	3	**140**	PF08644	2	**192**	PF01794	2	**244**	PF00169	2
37	PF03372	4	**89**	PF00132	3	**141**	PF08512	2	**193**	PF01740	2	**245**	PF00130	2
38	PF03105	4	**90**	PF00063	3	**142**	PF08506	2	**194**	PF01734	2	**246**	PF00111	2
39	PF02259	4	**91**	PF00027	3	**143**	PF08389	2	**195**	PF01645	2	**247**	PF00109	2
40	PF02134	4	**92**	PF15411	2	**144**	PF08326	2	**196**	PF01612	2	**248**	PF00097	2
41	PF00627	4	**93**	PF14878	2	**145**	PF08324	2	**197**	PF01593	2	**249**	PF00072	2
42	PF00581	4	**94**	PF14844	2	**146**	PF08321	2	**198**	PF01590	2	**250**	PF00036	2
43	PF00515	4	**95**	PF14691	2	**147**	PF08022	2	**199**	PF01571	2	**251**	PF00024	2
44	PF00512	4	**96**	PF14666	2	**148**	PF07724	2	**200**	PF01493	2	**252**	DUF913	2
45	PF00355	4	**97**	PF14664	2	**149**	PF07718	2	**201**	PF01422	2	**253**	DUF4414	2
46	PF00320	4	**98**	PF14663	2	**150**	PF07529	2	**202**	PF01315	2	**254**	DUF4187	2
47	PF00291	4	**99**	PF14641	2	**151**	PF07524	2	**203**	PF01266	2	**255**	DUF3385	2
48	PF00168	4	**100**	PF14639	2	**152**	PF07250	2	**204**	PF01243	2	**256**	DUF1929	2
49	PF00082	4	**101**	PF14635	2	**153**	PF06463	2	**205**	PF01119	2	**257**	DUF1771	2
50	PF00023	4	**102**	PF14624	2	**154**	PF05773	2	**206**	PF01096	2	**258**	DUF1752	2
51	PF13639	3	**103**	PF14604	2	**155**	PF05406	2	**207**	PF01077	2	**259**	DUF1720	2
52	PF13499	3	**104**	PF14598	2	**156**	PF05237	2	**208**	PF01068	2	**260**		

Where CC – connected component

5.4.7 The community structure detection

This analysis concentrates on the largest connected components of the domain-association network (Figure 5-15). The community structure of the main component was identified by means of the greedy agglomerative algorithm knows as Louvain method (Vincent et al., 2008). As a result, 25 communities with modularity equal to 0.8784 were detected (Figure 5-19).

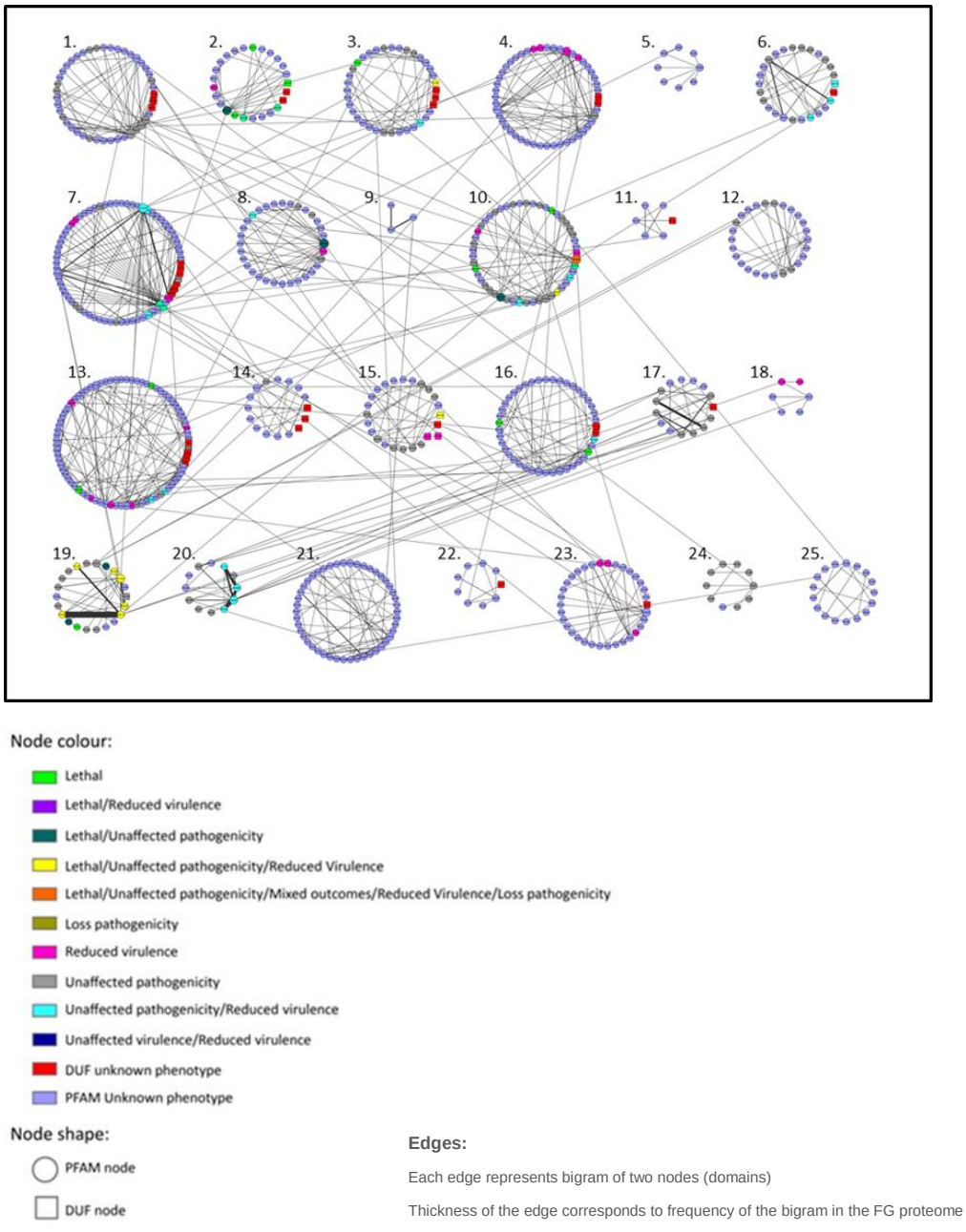


Figure 5-19 The community structure of the main component.
DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either it is a DUF (square) node or pfam node (circle).

The resultant community structure was found to have an uneven community-size distribution. This is illustrated in Figure 5-20. The number of DUF nodes within the modules is not proportional to the number of total nodes in the module. The DUF nodes were found to be present in only 16 out of 25 clusters.

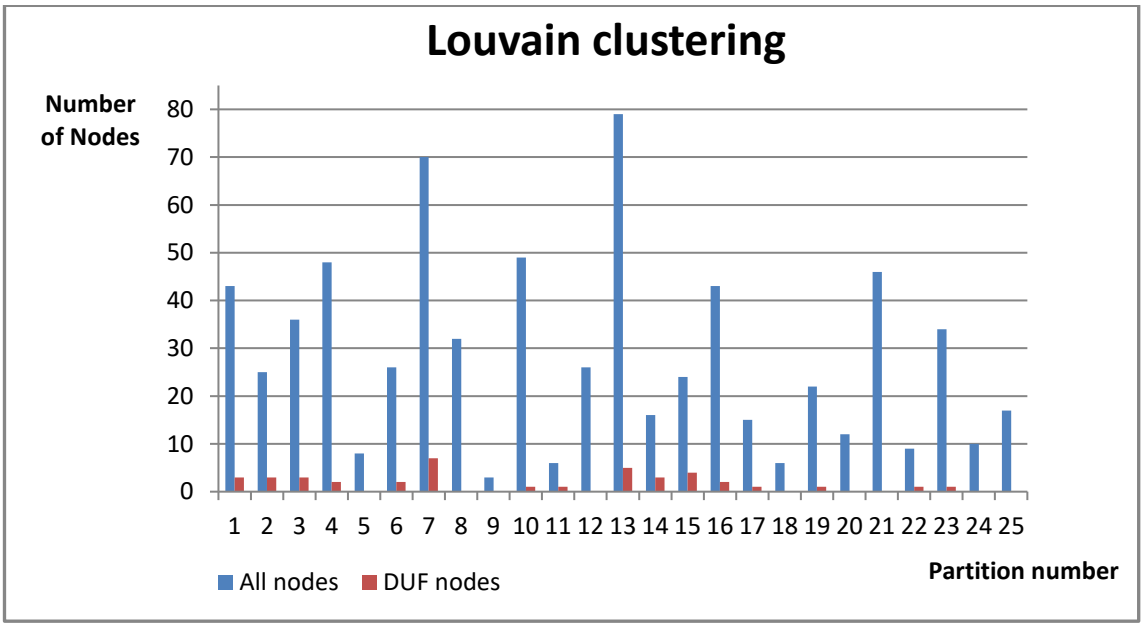


Figure 5-20 Modules size distribution of the main component of the domain-association network.

The majority of the communities revealed heterogeneous content in terms of domain phenotypes inherited from the proteins to which they belong. The highest discrepancy is observed in cluster 10 (Figure 5-21) which combines eight inherited phenotypes. The cluster consists of 49 unique pfam domains representing in a total of 199 FG proteins. The cluster also contains one DUF node (DUF3543), which is associated with reduced virulence protein FGSG_05547.

In addition, cluster 2 (Figure 5-22), cluster 7 (Figure 5-23), and cluster 19 (Figure 5-24) are also examples of the inconsistency within clusters in terms of inherited pfam domain phenotypes. Clusters: 2, 7 and 19 represent in total 30, 494, and 464 unique FG proteins respectively. Also, WoLF PSORT subcellular localisation prediction was performed on FG proteins represented by clusters 2, 7, 10, and 19 (with default setting: kNN = 27). The analysis indicates that the majority of FG proteins belonging to the above clusters through their pfam domains are intracellular proteins.

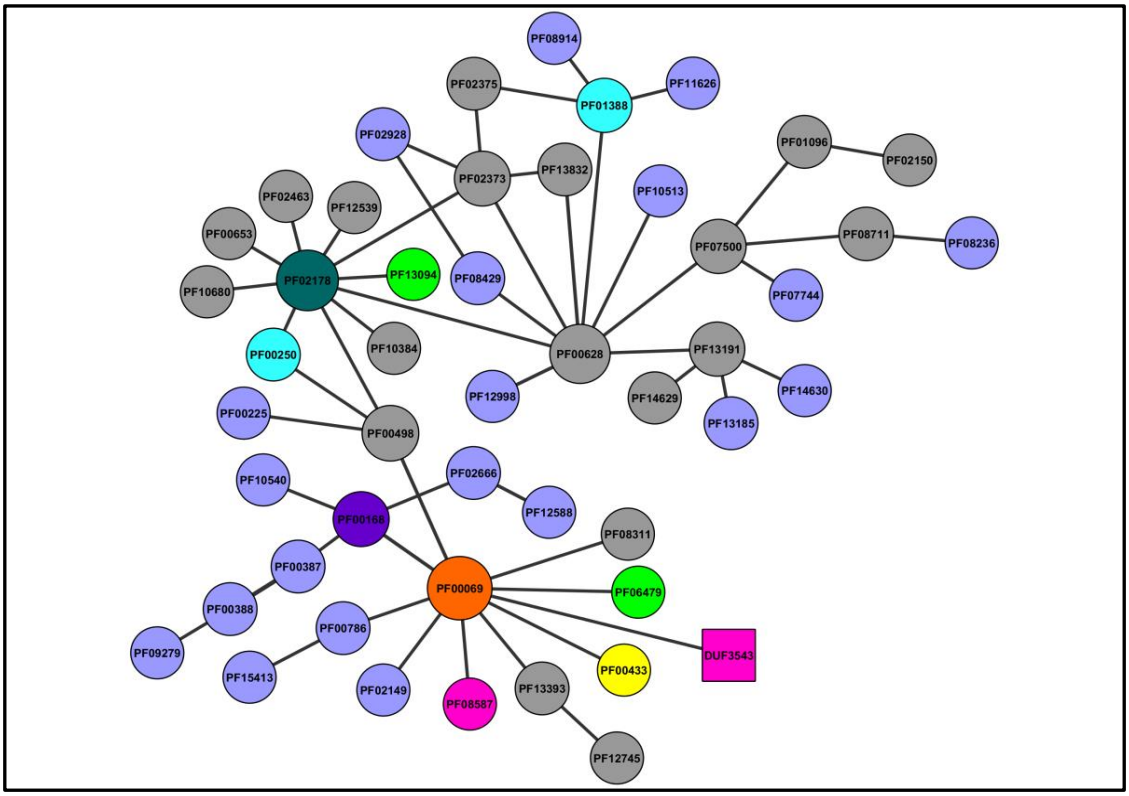
Moreover, cluster 15 (Figure 5-24) is an interesting example of the direct connection of the two domains with contrasting phenotypes (unaffected pathogenicity PF01585 node and reduced virulence DUF4187 node). Thus, based on the phenotype information within cluster 15 it is difficult to predict the phenotype of proteins FGSG_16727 and FGSG_05533 containing DUF1771 and protein FGSG_13541 containing DUF1604.

On the other hand, there are six homogenous communities with all domains derived from FG proteins with unknown phenotype. These are cluster 5, cluster 9, cluster 11, cluster 21, cluster 22, and cluster 25. Most of these clusters are very small ones with the number of nodes varying from 3 to 17. The exception here is cluster 21 which comprises 46 nodes. Two of the homogenous clusters with unknown phenotype have one DUF among the other pfam domains. These are cluster 11 and cluster 22 (Figure 5-19).

Moreover, there are six clusters with one phenotype identified within the given cluster. These are cluster 1, cluster 14, cluster 17, cluster 18, cluster 23, and cluster 24 (Figure 5-19). There are DUFs in cluster 1, cluster 14, cluster 17, and cluster 23. While inspecting cluster 14 and cluster 23 (Figure 5-25) it is difficult to speculate the possible phenotype of DUFs within the clusters as the phenotype is only assigned to 1 out of 16 pfam domain (PF13621) within cluster 14 and three out of 34 pfam domains (PF00481, PF13516 and PF13504) in cluster 23 and domains with the associated phenotype are neither the 1st nor 2nd neighbour node of the DUF nodes within these clusters.

However, whilst investigating cluster 1 and cluster 17 (Figure 5-26), where the majority of nodes were inherited unaffected pathogenicity phenotype, we might speculate that FG proteins: FGSG_05322 and FGSG_11656 that contain DUF1729, and FG proteins: FGSG_10896, FGSG_09740, FGSG_04350 and FGSG_05687 that contain DUF4217, as well as FGSG_07102 protein with DUF1998 might have no effect on the pathogenicity of FG.

In addition, cluster 24 (Figure 5-26) combines pfam domains associated with unaffected pathogenicity. The exception here is one node PF08292 with unknown phenotype. Thus, based on the information from cluster 24, we might speculate that protein FGSG_05602, containing PF08292 domain, is not involved in the pathogenicity process of FG.



Node colour:

- Lethal
- Lethal/Reduced virulence
- Lethal/Unaffected pathogenicity
- Lethal/Unaffected pathogenicity/Reduced Virulence
- Lethal/Unaffected pathogenicity/Mixed outcomes/Reduced Virulence/Loss pathogenicity
- Reduced virulence
- Unaffected pathogenicity
- Unaffected pathogenicity/Reduced virulence
- PFAM Unknown phenotype

Node shape:

- PFAM node
- DUF node

Edges:

Each edge represents bigram of two nodes (domains)

Thickness of the edge corresponds to frequency of the bigram in the FG proteome

Figure 5-21 Detailed content of cluster 10.
The figure illustrates in detail cluster 10 from Figure 5-19. DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either it is a DUF (square) node or pfam node (circle).

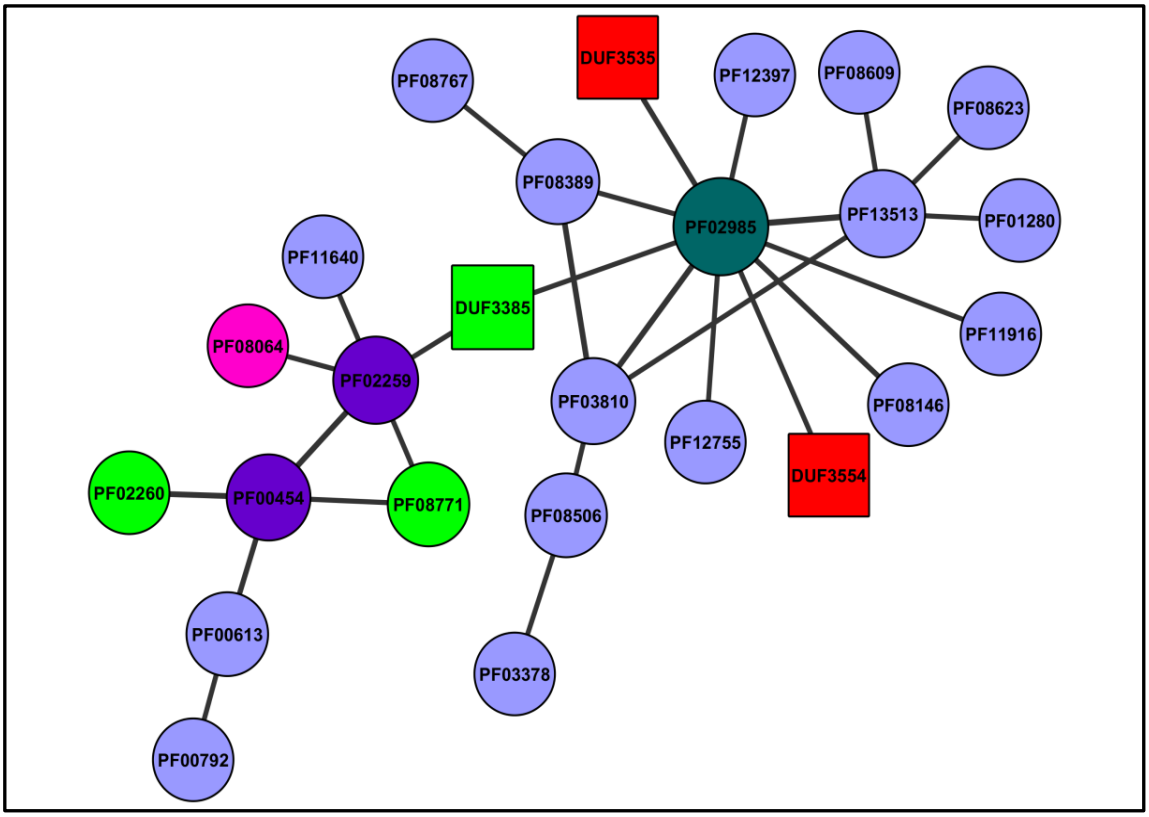
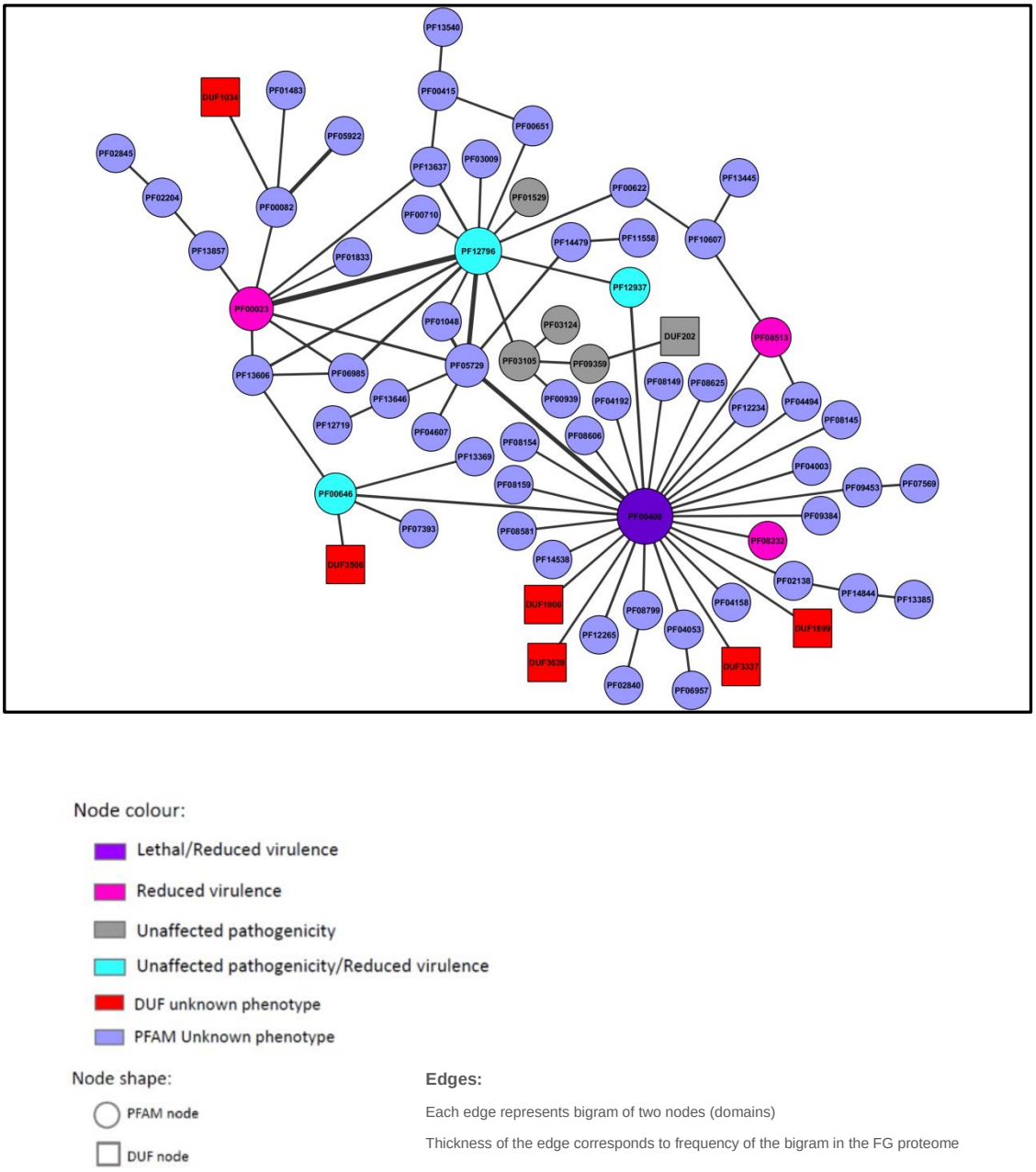


Figure 5-22 Detailed content of cluster 2.
The figure illustrates in detail cluster 2 from Figure 5-19. DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates that it is either a DUF (square) node or pfam node (circle).



Node colour:

- Lethal/Reduced virulence
- Reduced virulence
- Unaffected pathogenicity
- Unaffected pathogenicity/Reduced virulence
- DUF unknown phenotype
- PFAM Unknown phenotype

Node shape:

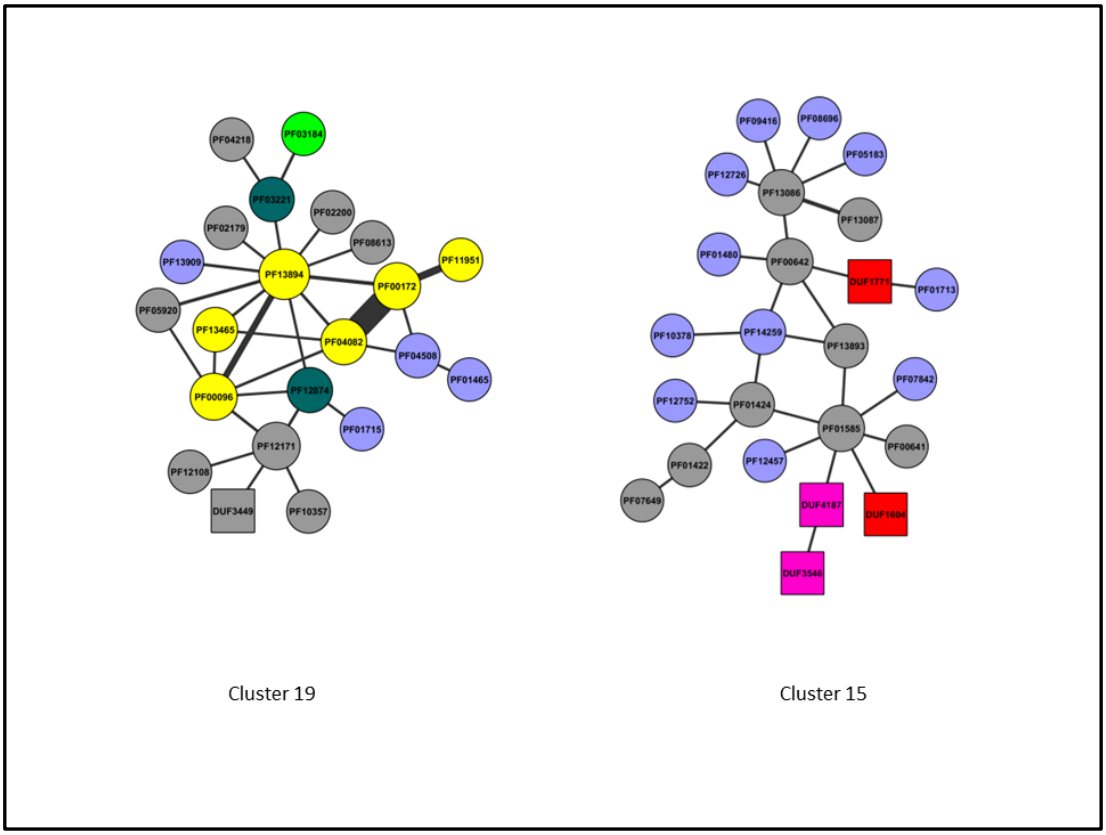
- PFAM node
- DUF node

Edges:

Each edge represents bigram of two nodes (domains)

Thickness of the edge corresponds to frequency of the bigram in the FG proteome

Figure 5-23 Detailed content of cluster 7.
The figure illustrates in detail cluster 7 from Figure 5-19. DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either that it is a DUF (square) node or pfam node (circle).



Node colour:

- Lethal
- Lethal/Unaffected pathogenicity
- Lethal/Unaffected pathogenicity/Reduced Virulence
- Reduced virulence
- Unaffected pathogenicity
- DUF unknown phenotype
- PFAM Unknown phenotype

Node shape:

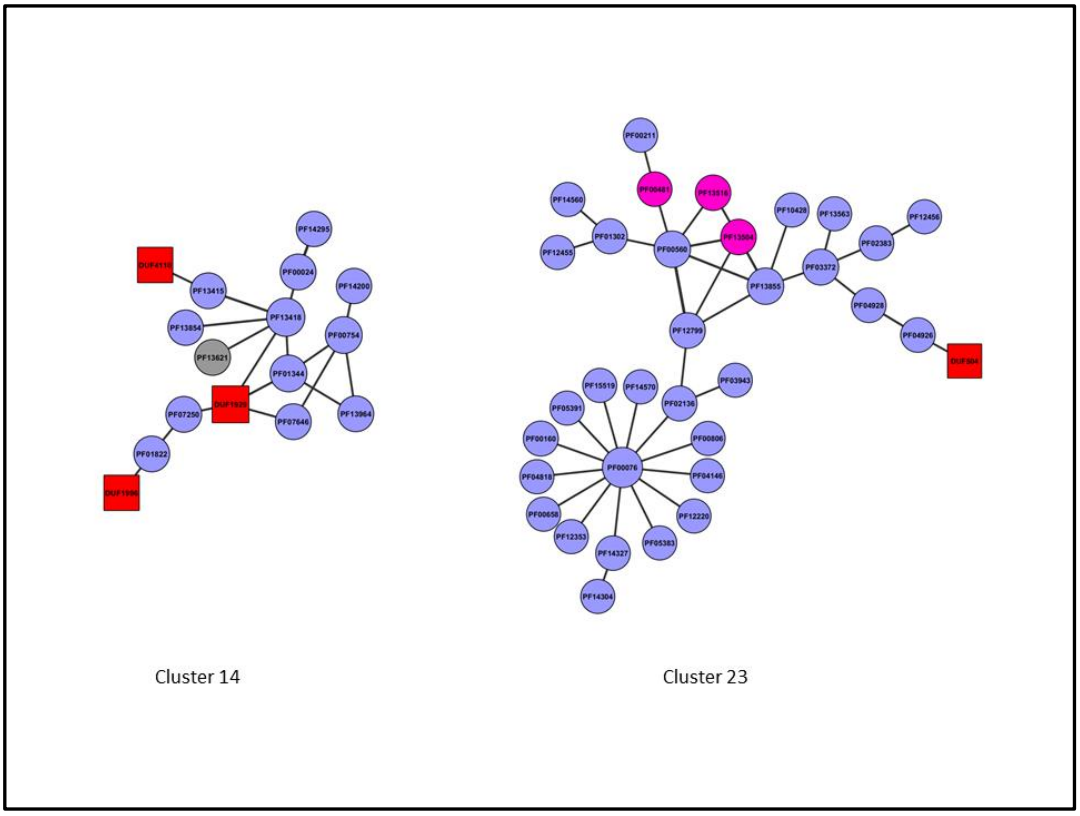
- PFAM node
- DUF node

Edges:

Each edge represents bigram of two nodes (domains)

Thickness of the edge corresponds to frequency of the bigram in the FG proteome

Figure 5-24 Detailed content of cluster 15 and cluster 19.
The figure illustrates in detail clusters 15 and 19 from Figure 5-19. DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either that it is a DUF (square) node or pfam node (circle).



Node colour:

- Reduced virulence
- Unaffected pathogenicity
- DUF unknown phenotype
- PFAM Unknown phenotype

Node shape:

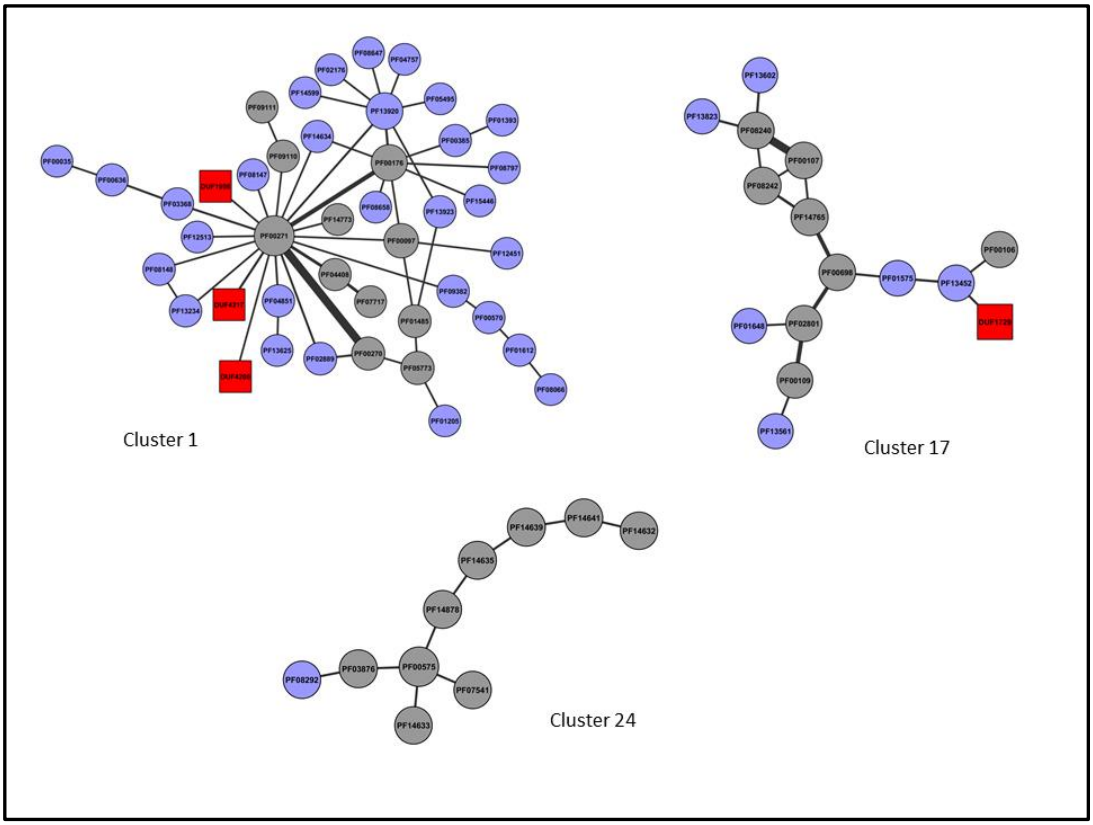
- PFAM node
- DUF node

Edges:

Each edge represents bigram of two nodes (domains)

Thickness of the edge corresponds to frequency of the bigram in the FG proteome

Figure 5-25 Detailed content of cluster 14 and cluster 23.
The figure illustrates in detail clusters 14 and 23 from Figure 5-19. DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either that it is a DUF (square) node or pfam node (circle).



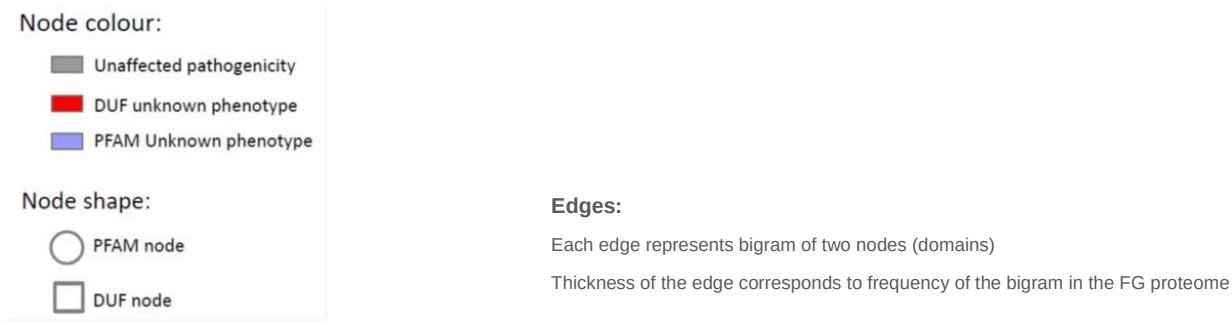


Figure 5-26 Detailed content of cluster 1, cluster 17 and cluster 24.
The figure illustrates in detail clusters 1, 17 and 24 from Figure 5-19. DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either that it is a DUF (square) node or pfam node (circle).

5.4.8 Functional cartography of clustered domains

In order to better describe the topological nature of the nodes that lie within the community structure detected in section 5.4.7, a node classification scheme has been employed. This approach offers a more advance description of the topological parameters of the node. The role of nodes, especially the DUF domains, within the community structure was determined based on the classification scheme proposed in (Guimera and Nunes Amaral, 2005). The distribution of node role types is recorded in Table 5-19 and the node distribution within the z-score and the participation coefficient space p is illustrated in Figure 5-27 generated by means of GIANT1.0 - Cytoscape 2.8.3 plugin.

Table 5-19 Node role distribution.

Role type	R1	R2	R3	R4	R5	R6	R7
Nodes No	600	61	29	3	10	2	0
Nodes No [%]	85.11	8.65	4.11	0.43	1.42	0.28	0

Where: R1 – ultra-peripheral node (all links within the cluster), R2 – peripheral node (most links within the cluster), R3 – non-hub connector node (many links to other clusters), R4 – non-hub kinless node (links homogeneously spread among all clusters), R5 – provincial hub (hub node with majority links within its cluster), R6 – connector hub (hub with many links to other clusters), R7 – global kinless hub (hub with links homogeneously spread among all clusters).

Overall, the majority of the nodes (85.11%) within the community structure are defined as ultra-peripheral nodes (R1) with all links within the cluster to which they belong. All DUF nodes belong to this group of nodes. This can suggest that the functionality of DUF nodes is strongly associated with the function of other nodes in the same cluster. There are no R7 nodes detected, which was observed in the previous study (Guimera and Nunes Amaral, 2005) to be common. There are only two nodes R6 (PF00400 and PF02985) defined as connector hubs with many links to most of the other clusters. Not surprisingly, these nodes were previously (section 5.4.6 of this chapter) found to be the articulation points of the main connected component of domain-association network and domain PF00400 is the most abundant domain not only within the main connected component of the network but in the whole network (please refer to Table 5-3 in section 5.4.1 of this chapter).

Whilst comparing the node associated phenotype to the node role, eight out of ten (80%) of lethal nodes present in the main connected component are ultra-peripheral nodes (R1) with their links solely within their module. One (10%) lethal node in the main connected component belongs to the peripheral node with most links within its module (R2) and one (10%) lethal node was classified as the non-hub connector node with many links to other modules (R3).

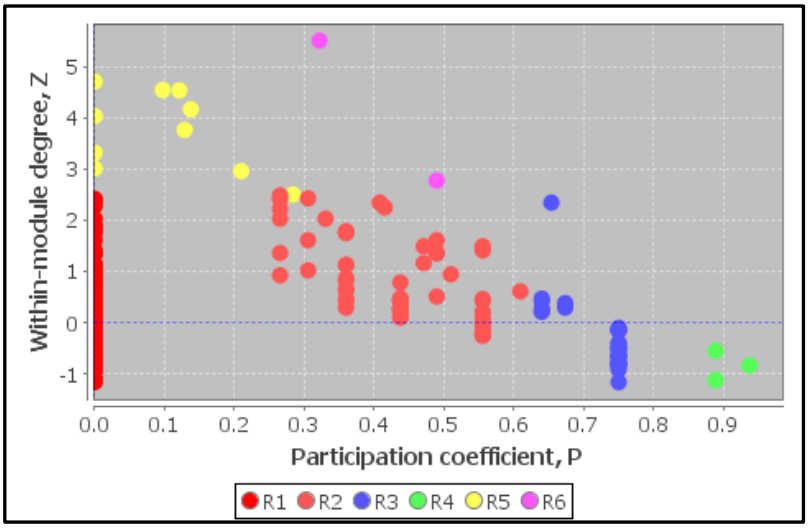


Figure 5-27 Nodes roles in the z-score and the participation coefficient space.
Where: R1 – ultra-peripheral node (all links within the cluster), R2 – peripheral node (most links within the cluster), R3 – non-hub connector node (many links to other clusters), R4 – non-hub kinless node (links homogeneously spread among all clusters), R5 – provincial hub (hub node with majority links within its cluster), R6 – connector hub (hub with many links to other clusters).

Furthermore, in reduced virulence nodes group, 15 out of 20 (75%) nodes are characterised as ultra-peripheral nodes (R1), four out of 20 (20%) nodes were categorised as peripheral nodes (R2) and one out of 20 (5%) nodes was classified as the non-hub connector node (R3). No loss pathogenicity nodes were detected within the main connected component of the network. In contrast, 74 (73.27%), 18 (17.82%), 7 (6.93%), 2 (1.98%) and 1 (~1%) of 101 unaffected pathogenicity nodes within the main component were characterised as ultra-peripheral (R1), peripheral (R2), non-hub connector (R3), non-hub kinless (R4) and provincial hub (R5) nodes respectively. DUF nodes, as mention earlier, were classified as ultra-peripheral nodes (R1) with their links within the module. While, R1 nodes are very well represented in every group of nodes with associated phenotype, it is difficult to speculate the role of DUF nodes in terms of associated

phenotype. However, within unaffected pathogenicity nodes there are a small number of nodes characterised as R4 and R5 which distinguishes them from lethal and loss virulence nodes.

5.4.9 Functional annotation of modules generated by the Louvain method

Following functional cartography analysis performed in section 5.4.8 and identification of DUF nodes as ultra-peripheral nodes that are tightly connected with their specific communities, I was interested in suggesting the functional properties of communities containing DUF nodes.

Communities in the network are likely to indicate that nodes share some common biological property. Thus, assigning a functional coherence to the modules generated by the Louvain method can suggest functionality of the DUF nodes belonging to each specific cluster. Therefore, this study focuses on the communities (previously calculated via the Louvain method in section 5.4.7) within the main component of the domain-association network (Figure 5-19).

The AIC-MICA metric (Lysenko et al., 2011) was applied in order to determine the annotation of Louvain clusters in terms of GO functional role at three levels: biological process (BioP), molecular function (MolF) and cellular component (CellC). Here, the AIC-MICA test was used to scale the degree of the identity of domains (nodes) annotation in particular cluster.

Overall, 16 clusters (including 12 with DUFs, Table 5-20) were assigned a role from the BioP feature of GO, 19 (including 13 with DUFs, Table 5-21) were assigned a function from the MolF feature of GO and 7 (including 6 with DUFs, Table 5-22) were assigned a term from the CellC feature of GO.

Figure 5-28 illustrates the corresponding MICA BiolP features, together with the coverage percentage for the annotated communities. The majority of the communities revealed less homogenous functional content; hence a broader GO term had to be assigned. Despite the low overall percentage coverage at all levels of Gene Ontology among the majority of the communities, the information defined by the AIC-MICA test can be vital for defining the functional coherence of DUF domains.

Six clusters that contain DUF nodes were assigned the GO functional role at all three levels. These are clusters 1, 3, 4, 7, 10, and 13. Additionally, for each of the above clusters the WoLF PSORT subcellular localisation prediction was performed on FG proteins that contain DUFs identified within each cluster (Table 5-23). The analysis enhances the finding from AIC-MICA test in terms of GO functional role at CellC level.

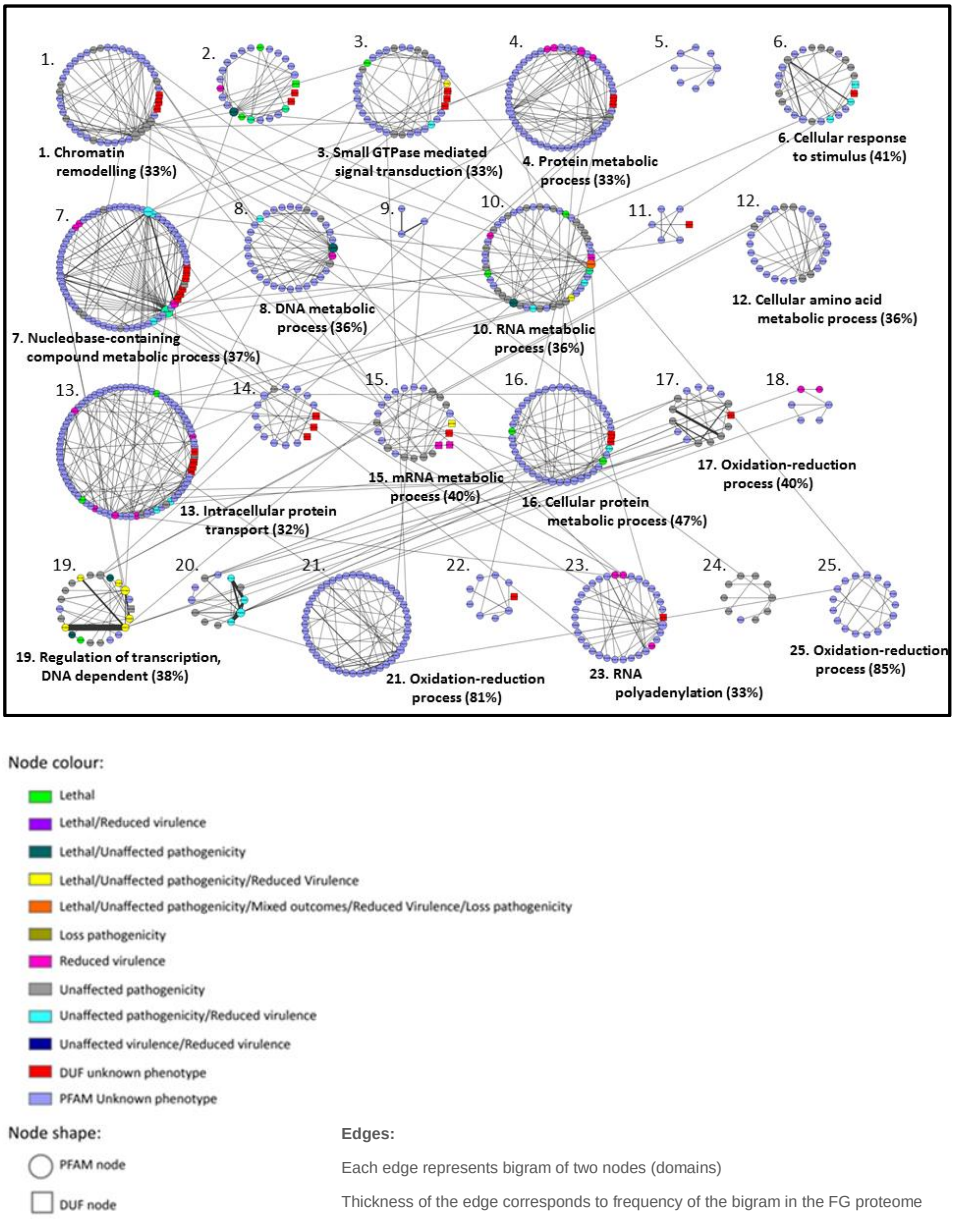


Figure 5-28 The community structure of the main component detected by Louvain method and annotated with the most informative GO biological process term at 30% threshold.
The MICA (BioP) term is shown beneath corresponding community with coverage percentage in the bracket. DUF domains with an unknown phenotype are highlighted in red. Different colours of the nodes indicate the associated phenotype or a group of phenotypes of *F. graminearum* protein (based on the information from PHI-base version 3.4) that contain the particular domain. The shape of the node indicates either that it is a DUF (square) node or pfam node (circle).

Table 5-20 Annotated communities with corresponding MICA biological process (BioP) terms.

Cluster Id	DUFs Id	AIC-MICA	GO Id	GO Name	Term IC	Coverage	Count
1	DUF4217 DUF4208 DUF1998	7.29	GO:0006396	RNA processing	5.28	0.33	2
		7.29	GO:0006338	**chromatin remodeling**	**9.30**	**0.33**	2
3	DUF908 DUF913 DUF4414	9.56	GO:0007264	small GTPase mediated signal transduction	9.56	0.33	2
4	DUF1977 DUF3395	2.76	GO:0019538	**protein metabolic process**	**3.28**	**0.33**	4
		2.76	GO:0044260	cellular macromolecule metabolic process	2.19	0.58	7
		2.76	GO:0090304	nucleic acid metabolic process	2.82	0.33	4
6	DUF1752 DUF3295	2.91	GO:0006355	regulation of transcription, DNA-dependent	3.85	0.35	6
		2.91	GO:0044260	cellular macromolecule metabolic process	2.19	0.41	7
		2.91	GO:0044238	primary metabolic process	1.46	0.41	7
		2.91	GO:0051716	**cellular response to stimulus**	**4.13**	**0.41**	7
7	DUF3506 DUF1899 DUF3337 DUF1034 DUF3639 DUF202	2.27	GO:0043170	macromolecule metabolic process	1.95	0.37	7
		2.27	GO:0006139	**nucleobase-containing compound metabolic process**	**2.59**	**0.37**	7
8		3.52	GO:0006259	**DNA metabolic process**	**3.75**	**0.36**	4
		3.52	GO:0019538	protein metabolic process	3.28	0.45	5
10	DUF3543	3.57	GO:0010467	gene expression	3.46	0.36	5
		3.57	GO:0016070	**RNA metabolic process**	**3.88**	**0.36**	5
		3.57	GO:0034645	cellular macromolecule biosynthetic process	3.37	0.36	5
12		4.80	GO:0006520	**cellular amino acid metabolic process**	**4.99**	**0.33**	4
		4.80	GO:0018130	heterocycle biosynthetic process	4.98	0.33	4
		4.80	GO:0044271	cellular nitrogen compound biosynthetic process	4.43	0.33	4
13	DUF21 DUF3694 DUF3608 DUF500	4.42	GO:0006886	**intracellular protein transport**	**5.93**	**0.32**	9
		4.42	GO:0050794	regulation of cellular process	2.91	0.32	9
15	DUF1604 DUF3546 DUF4187 DUF1771	5.37	GO:0006355	regulation of transcription, DNA-dependent	3.85	0.40	2
		5.37	GO:0016071	**mRNA metabolic process**	**6.88**	**0.40**	2
16	DUF3517 DUF1720	3.76	GO:0044267	cellular protein metabolic process	3.76	0.47	7
17	DUF1729	3.59	GO:0055114	oxidation-reduction process	3.59	0.40	2
19	DUF3449	2.38	GO:0006355	**regulation of transcription, DNA-dependent**	**3.85**	**0.38**	3
		2.38	GO:0009987	cellular process	0.91	0.38	3
21		3.59	GO:0055114	oxidation-reduction process	3.59	0.81	17
23	DUF504	6.73	GO:0006810	transport	3.15	0.33	2
		6.73	GO:0043631	**RNA polyadenylation**	**10.30**	**0.33**	2
25		2.25	GO:0009987	cellular process	0.91	0.31	4
		2.25	GO:0055114	**oxidation-reduction process**	**3.59**	**0.85**	11

Where AIC-MICA - The Average Information Content of the Most Informative Common Ancestor; DUFs – Domains of Unknown Function. When several GO annotations are available for a given cluster, the one with the highest Information Content (IC) term was highlighted in bold

Table 5-21 Annotated communities with corresponding MICA molecular function (MolF) terms.

Cluster Id	DUFs Id	AIC-MICA	GO Id	GO Name	Term IC	Coverage	Count
1	DUF4217 DUF4208 DUF1998	4.21	GO:0005524	ATP binding	4.40	0.32	7
		4.21	GO:0003676	nucleic acid binding	2.97	0.45	10
		4.21	GO:0016818	**hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides**	**5.25**	**0.36**	8
2	DUF3535 DUF3554 DUF3385	4.55	GO:0016773	**phosphotransferase activity, alcohol group as acceptor**	**5.47**	**0.30**	3
		4.55	GO:0005515	protein binding	3.63	0.50	5
3	DUF908 DUF913 DUF4414	3.63	GO:0005515	protein binding	3.63	0.35	7
4	DUF1977 DUF3395	3.12	GO:0016787	hydrolase activity	2.61	0.33	5
		3.12	GO:0005515	**protein binding**	**3.63**	**0.33**	5
6	DUF1752 DUF3295	4.64	GO:0003677	DNA binding	3.52	0.38	6
		4.64	GO:0005524	ATP binding	4.40	0.38	6
		4.64	GO:0004871	**signal transducer activity**	**6.00**	**0.31**	5
7	DUF3506 DUF1899 DUF3337 DUF1034 DUF3639 DUF202 DUF1900	3.63	GO:0005515	protein binding	3.63	0.56	10
8		2.77	GO:0003824	catalytic activity	1.15	0.65	11
		2.77	GO:0005524	**ATP binding**	**4.40**	**0.41**	7
10	DUF3543	2.39	GO:0003824	catalytic activity	1.15	0.48	10
		2.39	GO:0005515	**protein binding**	**3.63**	**0.33**	7
12		4.24	GO:0016740	transferase activity	2.90	0.31	5
		4.24	GO:0016874	**ligase activity**	**5.59**	**0.38**	6
13	DUF21 DUF3694 DUF3608 DUF500 DUF3818	1.25	GO:0003824	catalytic activity	1.15	0.32	9
		1.25	GO:0005488	**binding**	**1.34**	**0.43**	12
15	DUF1771 DUF4187 DUF1604 DUF3546	3.16	GO:0003824	catalytic activity	1.15	0.33	3
		3.16	GO:0008270	**zinc ion binding**	**5.34**	**0.33**	3
		3.16	GO:0003676	nucleic acid binding	2.97	0.44	4
16	DUF3517 DUF1720	1.25	GO:0003824	catalytic activity	1.15	0.50	10
		1.25	GO:0005488	**binding**	**1.34**	**0.50**	10
17	DUF1729	3.79	GO:0046872	metal ion binding	3.75	0.40	2
		3.79	GO:0016491	**oxidoreductase activity**	**3.83**	**0.80**	4
19	DUF3449	2.97	GO:0003676	nucleic acid binding	2.97	0.36	4
20		5.53	GO:0008168	methyltransferase activity	5.53	0.40	2
21		2.59	GO:0005488	binding	1.34	0.52	16
		2.59	GO:0016491	**oxidoreductase activity**	**3.83**	**0.58**	18
23	DUF504	3.04	GO:0003824	catalytic activity	1.15	0.60	6
		3.04	GO:0003723	**RNA binding**	**4.92**	**0.30**	3
24		4.92	GO:0003723	RNA binding	4.92	0.40	2
25		3.83	GO:0016491	oxidoreductase activity	3.83	0.91	10

Where AIC-MICA - The Average Information Content of the Most Informative Common Ancestor; DUFs – Domains of Unknown Function. When several GO annotations are available for a given cluster, the one with the highest Information Content (IC) term was highlighted in bold

Table 5-22 Annotated communities with corresponding MICA cellular component (CellC) terms.

Cluster Id	DUFs	AIC-MICA	GO Id	GO Name	Term IC	Coverage	Count
1	DUF4217 DUF4208 DUF1998	2.84	GO:0005634	nucleus	2.84	0.80	4
3	DUF908 DUF913 DUF4414	1.09	GO:0005622	intracellular	1.09	0.80	4
4	DUF1977 DUF3395	5.41	GO:0005634	nucleus	2.84	0.56	5
		5.41	GO:0000151	**ubiquitin ligase complex**	**7.99**	**0.33**	**3**
7	DUF3506 DUF1899 DUF3337 DUF1034 DUF3639 DUF202 DUF1900	2.20	GO:0032991	macromolecular complex	2.35	0.38	5
		2.20	GO:0005634	**nucleus**	**2.84**	**0.46**	6
		2.20	GO:0016020	membrane	1.40	0.31	4
10	DUF3543	4.11	GO:0005634	nucleus	2.84	0.67	4
		4.11	GO:0005694	**chromosome**	**5.38**	**0.33**	2
13	DUF21 DUF3694 DUF3608 DUF500 DUF3818	4.41	GO:0043229	intracellular organelle	1.76	0.33	5
		4.41	GO:0030117	**membrane coat**	**7.07**	**0.33**	**5**
21		9.16	GO:0016459	myosin complex	9.16	0.50	3

Where AIC-MICA - The Average Information Content of the Most Informative Common Ancestor; DUFs – Domains of Unknown Function. When several GO annotations are available for a given cluster, the one with the highest Information Content (IC) term was highlighted in bold

Table 5-23 WoLF PSORT subcellular localisation prediction for DUFs identified in annotated communities with corresponding MICA cellular compartment (CellC) terms.

Cluster id	DUFs id	FGSG id	WoLF PSORT subcellular localisation prediction
1	DUF4217	FGSG_10896	nucl: 10, cyto: 9, mito: 8
		FGSG_09740	cyto: 12, cyto_mito: 10.833, cyto_nucl: 9.833, mito: 8.5, nucl: 6.5
		FGSG_04350	nucl: 24.5, cyto_nucl: 14
		FGSG_05687	nucl: 22.5, cyto_nucl: 13.5, cyto: 3.5
	DUF4208	FGSG_07102	nucl: 22, cyto: 3
	DUF1998	FGSG_12034	nucl: 17.5, cyto_nucl: 12.5, cyto: 6.5
3	DUF908	FGSG_00633	nucl: 12, cyto_nucl: 10, cyto: 6, mito: 4, plas: 4
	DUF913	FGSG_00634	nucl: 12, cyto_nucl: 10, cyto: 6, mito: 4, plas: 5
	DUF4414	FGSG_00635	nucl: 12, cyto_nucl: 10, cyto: 6, mito: 4, plas: 6
4	DUF1977	FGSG_01620	mito: 9, nucl: 5, cyto: 4, plas: 4, pero: 3
	DUF3395	FGSG_00939	mito: 12, nucl: 8, cyto_nucl: 7.5, cyto: 5
7	DUF3506	FGSG_07514	nucl: 15.5, cyto_nucl: 11.5, cyto: 6.5, pero: 4
	DUF1899	FGSG_13118	mito: 22, nucl: 3
	DUF3337	FGSG_04351	nucl: 16, mito: 7, cyto: 2
	DUF1034	FGSG_06572	extr: 23, plas: 2
		FGSG_11472	extr: 24, cyto: 3
	DUF3639	FGSG_10869	nucl: 14.5, cyto_nucl: 11, cyto: 6.5, mito: 3
	DUF202	FGSG_00502	mito: 9, plas: 7, nucl: 4, pero: 4, cyto: 3
		FGSG_01432	nucl: 10.5, plas: 8, cyto_nucl: 7, mito: 3, cyto: 2.5
		FGSG_09731	plas: 17, mito: 3, E.R.: 3, nucl: 2
		FGSG_06868	plas: 18, E.R.: 3, nucl: 2, mito: 1, cyto: 1
		FGSG_09952	plas: 14, E.R.: 7, vacu: 3, mito: 2
	DUF1900	FGSG_13118	mito: 22, nucl: 3
10	DUF3543	FGSG_05547	nucl: 12, cyto_nucl: 11.833, mito_nucl: 9.999, cyto: 8.5, mito: 6.5
13	DUF21	FGSG_09484	plas: 13, mito: 5, E.R.: 4, cyto: 3
	DUF3694	FGSG_10189	nucl: 15.5, cyto_nucl: 13, cyto: 9.5
	DUF3608	FGSG_13640	nucl: 17.5, cyto_nucl: 10.5, mito: 6
	DUF50**0**	FGSG_03563	plas: 7, mito: 6, nucl: 5, pero: 4, cyto: 2, E.R.: 1
		FGSG_02018	cyto: 20, cyto_nucl: 13.5, nucl: 5
		FGSG_09497	mito: 11, cyto: 7, nucl: 3, extr: 3, pero: 3
		FGSG_16553	cyto: 26
		FGSG_01702	cyto_nucl: 14.5, nucl: 13.5, cyto: 10.5
		FGSG_16612	nucl: 10, mito: 5, cyto: 5, plas: 3, golg: 2
	DUF3818	FGSG_02586	nucl: 13.5, cyto_nucl: 11.5, cyto: 8.5, cysk: 5
		FGSG_13137	cyto: 12, cysk: 9, pero: 3, nucl: 2

Where DUFs – Domains of Unknown Function.; cellular compartments are cyto – cytosol, cysk – cytoskeleton, E.R.- endoplasmic reticulum, extr - extracellular location, golg - Golgi apparatus, pero - peroxisome, plas - plasma membrane, mito – mitochondria, nucl – nucleus, vacu – vacuole.

5.5 Discussion

In this chapter, several studies were conducted to investigate the possible role of DUF in disease-causing ability of the globally important plant pathogenic fungus *Fusarium graminearum* (FG). A combined strategy, incorporating FG pfam domain-repertoire identification (HMMR with Pfam), FG pfam-domain taxonomic-diversity evaluation (combined information from PFAM database version 27.0 and UniProt database (Consortium, 2014)), adjacent-domain bigrams-recognition (method adopted from previous study (Seidl et al., 2011)), as well as the implementation of the network analysis (NetworkX Python package, Cytoscape), was shown to be collectively informative in exploration of DUFs characteristics, possible function(s), and a role in the lifestyle of FG.

About 61% of FG proteome has been annotated with at least one pfam domain, whereas many of these FG proteins have only one pfam domain and a small percentage of them (5.6%) are DUFs. The proteins with only one pfam domain that is a DUF are mainly hypothetical proteins with neither experimental nor computationally predicted function. The genes coding for this protein type were found to be evenly distributed throughout the four chromosomes of FG. While, the lower recombination regions are in the majority throughout the FG four chromosomes, large numbers of DUFs are present in conserved regions of the FG genome which are predicted to be less evolutionary adaptive compared to proteins occupying the higher recombination regions (Cuomo et al., 2007, Ma et al., 2010). This finding might suggest that proteins with a single domain that is DUF are essential proteins. A similar observation was reported by Goodacre et al. (2014) when investigating the DUFs in bacteria.

In addition, one FG protein, namely FGSG_01939 with only one domain that is a DUF (DUF619) was experimentally proven to be required for the virulence (PHI-base version 3.4, version 3.6) (Lysenko et al., 2013, Urban et al., 2015). This outcome might suggest that proteins with only one domain that is DUF could have a direct role in the pathogenic lifestyle of FG.

As mentioned earlier in the thesis, the role of each domain results from the synergistic relationship with other domains either in the same proteins or interacting proteins. The study in this chapter

identified numbers of bigrams, pairs, or adjacent pfam domains in FG proteins. Many identified domain pairs are hetero-bigrams. Of these, there are 20 unique bigrams containing only DUFs. Unsurprisingly, they are annotated as hypothetical proteins, but are mostly assigned to low recombination regions of the genome. They are also predominantly predicted to be intracellular proteins. One such protein, namely FGSG_01106 (DUF3546|DUF4187), was identified as required for the virulence (PHI-base version, 3.4 and 3.6 (Lysenko et al., 2013, Urban et al., 2015)) and it is predicted to be located in the nucleus. This finding provides further evidence which might suggest that DUFs can be important factors in the plant pathogen lifestyle.

The taxonomical study of pfam domains and DUFs within the FG proteome identified 35 DUFs that are fungal-specific. The majority of these DUFs are the only domain in the associated protein and are located within the lower recombination regions of FG genome. Surprisingly, a detailed analysis of 35 fungal-specific DUFs revealed that both pathogenic and non-pathogenic fungi share the same repertoire of these 35 DUFs. However, DUF3129 and DUF3517 were identified to be the most abundant in pathogenic fungal species. Moreover, nine DUFs, namely DUF1687, DUF2417, DUF3292, DUF3835, DUF3425, DUF4451, DUF3844, DUF3602 and DUF3779 were found to be highly enriched within plant pathogenic fungi. Both the most abundant DUFs in pathogenic fungal species and those highly enriched within plant pathogenic fungi account for the only domain within the given protein and are associated with the lower recombination region of FG genome. Additionally, the proteins comprising these DUFs lack experimentally verified phenotype information. The exception here is DUF3425 which is present in 12 proteins and one of them, namely FGSG_12345 is associated with an unaffected pathogenicity phenotype.

Additionally, the chi-square test revealed a stronger association of DUF3292 with plant pathogenic fungi, whereas a slightly weaker association of DUF4448 with non-pathogenic fungi were identified. Both domains are the only domain in the linked proteins and most of these proteins are within the low recombination region of FG genome with exception of FGSG_10498 (represented by DUF3292) which is linked to middle recombination region of FG genome.

Furthermore, the extended analysis of 35 DUFs specific to fungal species demonstrated that their whole repertoire is present in nine plant pathogens representing a variety of a different lifestyles.

However, this group of plant pathogens are all Ascomycota and classical hemi-biotrophs. Thus, there is no evidence that a DUF which is present in other kingdoms has been transferred only into FG.

The implementation of the domain-association network shed light on further properties of DUFs. Statistical comparison of topological properties of DUF nodes with pfam nodes associated with phenotype, revealed similarity in property distribution between DUF nodes and lethal nodes. Moreover, distribution of properties of nodes associated with loss of pathogenicity phenotype was found not to be significantly different to the distribution of properties associated with DUF nodes. However, due to the small sample size of nodes associated with the loss pathogenicity phenotype this outcome is not meaningful. On the other hand, a high discrepancy was observed when comparing the distributions of topological properties of DUF nodes to one of the pfam nodes associated with unaffected pathogenicity phenotype. This finding might suggest that DUFs are linked to FG essential proteins.

Application of the node classification scheme to the nodes that lie within identified communities of the largest connected component uncovered that all DUF nodes are non-hub ultra-peripheral (R1) nodes with all links within their cluster. While, R1 nodes are also well represented among pfam nodes associated with different phenotypes, nodes characterised as R4 (non-hub nodes with links equally spread among all clusters) and R5 (hub-node with majority links within its cluster) were only identified with unaffected pathogenicity pfam nodes. This outcome distinguishes nodes connected with unaffected pathogenicity from nodes associated with lethal and loss virulence phenotypes.

Overall, the outcomes of studies conducted in this chapter strongly suggest that DUFs are highly distributed within the conserved proteins throughout the entire FG genome. Additionally, many DUFs are in single-domain proteins. Similar to a study of DUFs in the bacterial kingdom (Goodacre et al., 2014), DUFs in FG are part of the essential protein set. This means one of two things. One option is that the protein without which the organism would not exist and as such is needed for basic organism functions. This is very well supported by a study of the similarity in the

distribution of network node properties of DUFs and lethal nodes. On the other hand, the protein may be required for the pathogenic lifestyle of FG.

As pathogenicity is a complex process that requires synergistic relationship among several domains either in the same or among interacting proteins, DUF proteins can play an important part in the mechanism of the pathogenicity process. DUFs were shown to be peripheral rather than core domains in the domain-association network and the most of them were spread across small connecting components. This further suggests that DUFs are at least versatile domains which can appear together with other domain partners within the protein but the number of partner domains for the specific DUF is limited.

Furthermore, identifying nine hemibiotrophic Ascomycota that share the whole repertoire of 35 fungi-specific DUFs with FG also indicates that DUFs are part of the essential protein set for the particular Phylum. Although the role of DUFs in the pathogenicity of the plant pathogenic fungus FG is still unclear, a more in-depth study of the 71 FG proteins representing 35 fungi-specific DUFs might shed light on further properties of DUFs in FG. Also, the comparison between *Fusarium graminearum* and recently sequenced plant pathogen *Fusarium culmorum* and newly sequenced non-pathogenic *Fusarium venenatum* could possibly reveal the difference in DUF comparison between plant pathogens and non-pathogens within the *Fusarium* genus. This analysis is addressed in Chapter 6.

Chapter 6

DUF comparison amongst *Fusarium* genus members

6.1 Introduction

Following the successful completion of DUF analysis for *Fusarium graminearum* (Chapter 5), it was then considered interesting to explore and compare DUF composition in further *Fusarium* genus members including both plant pathogen and non-pathogen examples. Therefore, the aim of this chapter is to investigate whether there is a discrepancy between DUF frequencies of plant and non-plant pathogenic fungi of the same genus. For that purpose, comparative analysis of DUF repertoires was performed among three *Fusarium* members: *Fusarium graminearum* (FG), *Fusarium culmorum* (FC) and *Fusarium venenatum* (FV). Both FG and FC are plant pathogens on several small-grain cereals, especially wheat and barley, leading not only to significant grain yield lost but also to grain contamination with harmful mycotoxins (See section 2.2.2 of Chapter 2). In contrast, FV is a fungus of the genus *Fusarium* with non-pathogenic lifestyle. It has high protein content and one of its strains, namely A3/5 was developed commercially to obtain proteins (Quorn™) used as a substitution of meat diet.

6.2 Resources and methods

Protein sequences for FC and FV were downloaded from Rothamsted Research internal resources. In addition, as a new annotation of *F. graminearum* (FGRRES) genome became available at the time of writing this chapter (King et al., 2015), the proteins of the new improved version of the FG genome were downloaded from Ensembl fungi (Kersey et al., 2014). The pfam-domain repertoires (including DUFs) of FGRRES, FC and FV were identified using the same methodology and PFAM database version 27.0 as per Chapter 5. The overlapping issues were also solved by applying the custom pipeline implemented in Chapter 5 (Figure 5-2).

Furthermore, several DUFs detected in this study were examined in detail using information from the current (at the time of writing Chapter 6) PFAM database version 28.0. Also, a BLASTP protein

similarity search against NCBI nr was performed on protein / genes containing DUFs of interest. In this search default BLASTP algorithm parameters were used as described on NCBI BLASTP search.[35]

6.3 Results

6.3.1 General overview of DUFs across *Fusarium* proteomes

Table 6-1 Basic statistics of the different reference and annotation.

Features	FGRRES	FC	FV
Genome size (bp)*	36,570,348	39,005,997	38,580,615
Unknown bases (N)	12	2,922,878	859
Scaffolds	5	6	5
Chromosomes	4	4	4
Telomeres at chromosome ends	4	1	4
GC (%) content	48.2	47.6	47.7
Predicted Genes	14,164	13,929	13,946
ENA project accession	PRJEB5475	N/A	N/A

*Excluding N's, mitochondria and large repetitive sequence at the carboxyl end of chromosome IV.
FGRRES – *Fusarium graminearum* Rothamsted Research genome assembly, FC- *Fusarium culmorum* and FV – *Fusarium venenatum.*

Table 6-1 highlights basic genome statistics of the three recently annotated *Fusarium* genomes of interest in this study, whereas Table 6-2 reveals proteomic properties of these fungi. The 3Mb of unknown bases in the FC genome is an outcome of very fragmented FC genome draft assembly available at the time of the analysis performed in this chapter. As a consequence of this, we observe a significantly smaller predicted proteome compared with FGRRES and consequently a lower number of proteins annotated with pfam domains in FC compared to FGRRES, FG and FV (Table 6-2).

[35] http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

The abundance of each DUF within the predicted protein repertoire of FGRRES (n= 14,164), FC (n = 13,929) and FV (n = 13,946) were calculated and compared to FG (n = 13,826). Table 6-2 summarises the basic statistics of pfam domains and DUFs comparison amongst four proteomes. Overall, in all four proteomes DUFs represent the only a small percentage of the total pfam repertoire accounting for the highest number in FC (4.05%), where the number of proteins with at least one pfam domain / or DUF is the lowest. This is very well noticeable while comparing FC and FV, which have proteomes similar in size. This finding confirms that the quality of the FC assembly is not as good as those of FV and FGRRES. Furthermore, the majority of DUFs occur as a single copy throughout the total proteome, ranging from 60% in FV to 62% in FG.

Investigating further Table 6-2, it can be observed that DUF content slightly varies between the different species within the *Fusarium* genus being the lowest for FG and highest for FV. The same tendency is observed while comparing the number of unique DUFs across these species. However, while comparing the number of proteins with only one domain that is DUF, the small inconsistency to the previous trend is observed in FC. This finding might be an outcome of a very fragmented FC genome assembly with the indicative 3Mb of unknown bases (Table 6-1).

Table 6-2 DUF comparison amongst four *Fusarium* proteomes.

Species	FG	FGRRES	FC	FV
Proteome size (n)	13,826	14,164	13,929	13,946
Proteins with pfam domain(s)*	8,478	8,571	7,927	8,831
Total pfam domains**	13,217	13,402	12,961	13,653
Total DUFs number	504	515	525	538
Unique pfam domains Ids**	3,524	3,536	3,496	3,539
Unique DUFs Ids	314	319	320	326
One DUF only proteins number***	338	342	308	357
Total DUF number / Total pfam number [%]	3.81	3.84	4.05	3.94
Unique DUFs number / Total DUFs number [%]	62.30	61.94	60.95	60.59
Unique ordered bigrams number	1730	1750	2214	1789
Unique ordered bigrams with at least one DUF	111	112	161	118
Unique ordered DUF only bigrams	20	20	21	20

Where FG is *F. graminearum* FG3 MIPS gene call, FGRRES is *F. graminearum* Rothamsted Research gene call, FC is *F. culmorum* and FV is *F. venenatum*; *number of proteins with at least one pfam domain, including DUF. **pfam domains Ids include DUFs, ***number of proteins that contain one only pfam domain that is DUF.

6.3.2 Comparison of the most abundant DUFs

Table 6-3 summarises the most abundant DUFs amongst four proteomes (including two genome assemblies for *F. graminearum*, namely FG and FGRRES), whereas Figure 6-1 illustrates the distribution of these DUFs for each proteome. It is interesting to note that some DUFs are more abundant in the FV compared to FG, FGRRES and FC. These are DUF3433, DUF3425, DUF3659, DUF4267, DUF4066 and DUF1996. Both the most abundant DUFs in FV, namely DUF3433 and DUF3425, were also identified as specific to fungal species in Chapter 5 (see section 5.4.3 of Chapter 5) and both were also found to be more abundant within pathogenic fungi and fungal plant pathogens (see Figure 5-11 and 5-12 in Chapter 5). Thus, it is not understood why in this analysis the abundance of these DUFs is higher in the non-pathogenic FV proteome. The difference in abundance between pathogenic and non-pathogenic fungi is well pronounced for DUF3433 (Table 6-3).

As it was shown in Chapter 5 Table 5-7, DUF3433 commonly appears as the only homo-bigram in seven FG proteins and as a single domain in two FG proteins: FGSG_04690 and FGSG_00063; or their corresponding proteins in FGRRES gene call: FGRRES_04689_M and FGRRES_00063 respectively. Then again in the FC proteome, DUF3433 appears as the only homo-bigram in six FC proteins and in one protein, namely FCUL_08350, as one homo-bigram with the neighbourhood of the PF02811 domain. The latter is a putative phosphoesterase domain (Finn et al., 2014a). In FV proteome, however, DUF3433 is part of the only homo-bigram present in 22 proteins and a single-domain in two proteins (FV_07945 and FV_10435).

On the contrary, DUF3425 largely appears as a single domain in proteins. The exceptions are FGSG_12345/FGRRES_12354 (linked to the pfam domain PF00170a bZIP transcription factor), FCUL_11640 (linked to PF00106, dehdrogenases), FV_06884 (linked to PF00172) and FV_06981 (linked to PF00170). Moreover, as it was reported earlier in Chapter 5, protein FGSG_12345 is associated with an unaffected pathogenicity phenotype in PHI-base version 3.6 and the current PHI-base version 3.8 (Urban et al., 2015b). As reported earlier in Chapter 5 table 5-3, PF00172 is a fungal-specific domain.

Further examination of Table 6-3 revealed that some of DUFs are more common within pathogenic fungi proteomes compared to the non-pathogenic FV. These are DUF2235, DUF3632, DUF3129, DUF1752 and DUF1275. Although the difference in the abundance is very small, it is worth underlining that DUF3129 was also identified as the one associated with plant pathogenic fungi (see table 5-12).

Table 6-3 The most abundant DUFs across four *Fusarium* proteomes.

No	DUF Id	FG	FGRRES	FC	FV
1	**DUF3433**	16	17	14	24
2	**DUF3425**	12	13	13	16
3	**DUF3659**	11	11	7	12
4	**DUF2235**	8	9	9	8
5	**DUF500**	6	6	6	6
6	**DUF4267**	6	6	6	7
7	**DUF4463**	5	5	8	5
8	**DUF4066**	5	5	5	6
9	**DUF3328**	5	5	4	4
10	**DUF3237**	5	5	5	5
11	**DUF202**	5	5	6	5
12	**DUF1929**	5	5	5	5
13	**DUF829**	4	4	4	4
14	**DUF4217**	4	4	4	4
15	**DUF3632**	4	4	4	2
16	**DUF3602**	4	4	4	4
17	**DUF3129**	4	4	4	3
18	**DUF221**	4	5	5	5
19	**DUF1996**	4	4	4	5
20	**DUF1752**	4	4	4	3
21	**DUF1275**	4	4	4	3

Where FG and FGRRES are different calls of *F. graminearum* genome assembly: FG3 MIPS and Rothamsted Research respectively; FC - *F. culmorum* and FV – *F. venenatum*.

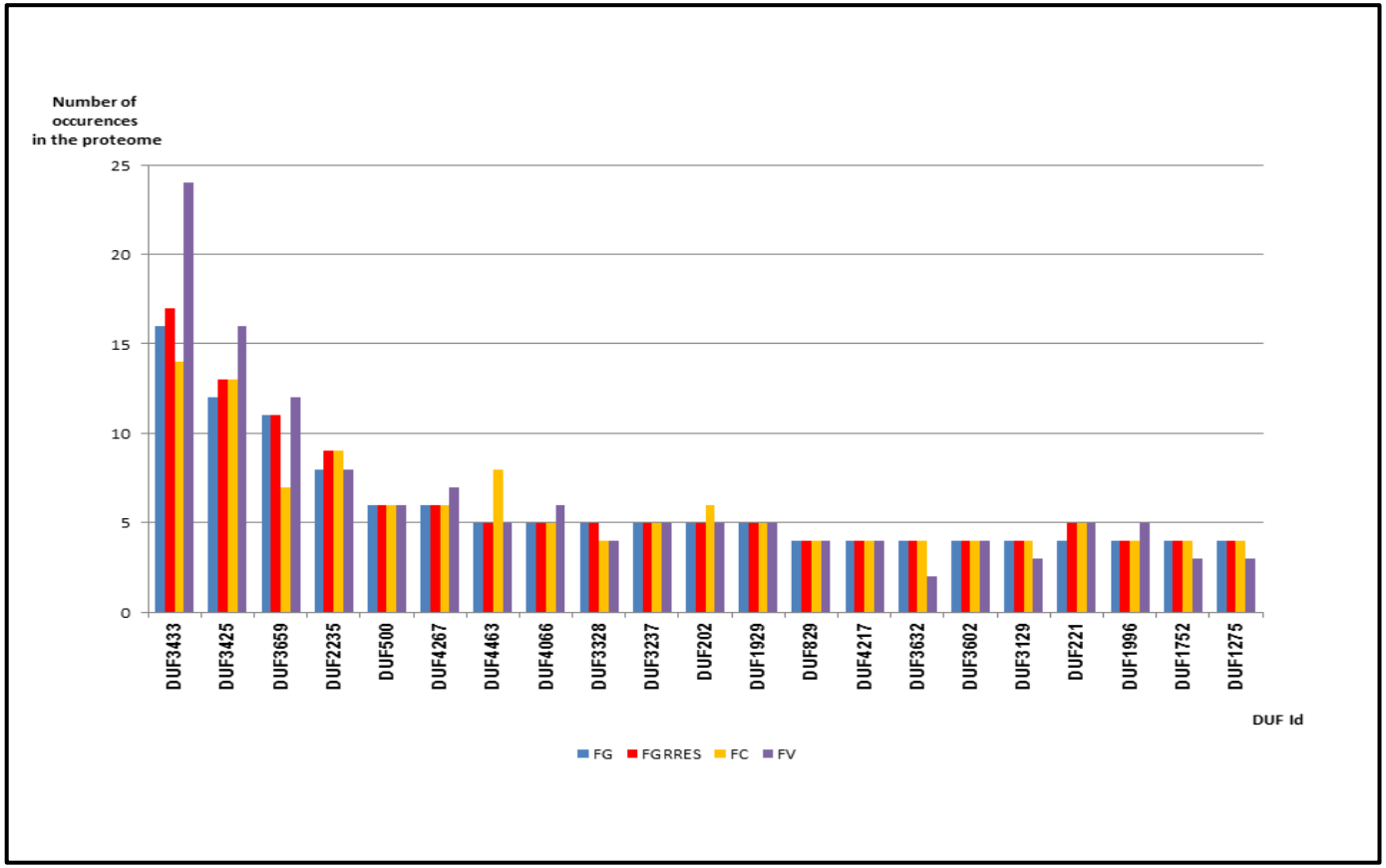


Figure 6-1 Distribution of the most abundant DUFs across four *Fusarium* proteomes.
Where FG – *F. graminearum* FG3 MIPS gene call, FGRRES - Rothamsted Research *F. graminearum* gene call, FC – *F. culmorum* and FV – *F. venenatum*.

6.3.3 Distribution of DUFs within tested *Fusarium* proteomes

In total, there are 330 unique DUFs identified within four *Fusarium* proteomes in this study. As seen from Figure 6-2, 308 DUFs are common to all four proteomes. Furthermore, one DUF, namely DUF1330 was only found in *F. graminearum*, see Table 6-4, whereas DUF1903 was only found in fungal pathogens: *F. graminearum* (both gene calls) and *F. culmorum*. However, DUF1903 has been renamed to Mature T-Cell Proliferation I type cysteine motif (MTCP1) in the PFAM database version 28.0 (Finn et al., 2014a).

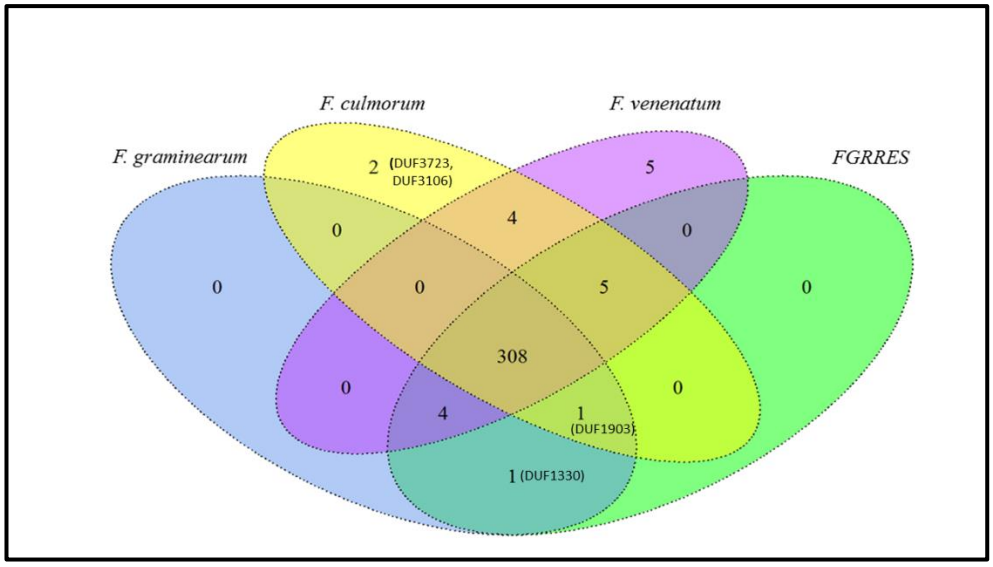


Figure 6-2 Distribution of DUFs within four *Fusarium* proteomes.
Where FGRRES – is a Rothamsted Research *F. graminearum* genome assembly, whereas *F. graminearum* set indicates FG3 MIPS genome assembly.

Further investigation of Table 6-4 indicates that DUF1330 exists as the only domain of FGSG_11455/FGRRES_11455. Furthermore, a BLASTP search against current NCBI nr database shows that this protein is *F. graminearum* specific. Moreover, additional examination of DUF1330 using PFAM version 28 revealed that this domain is very common in bacteria and only found in two fungal species: *F. oxysporum* and *F. graminearum*.

Also, four DUFs in common for FG, FGRRES and FV but not FC were identified (see Table E-1, Appendix E). They are single-domain proteins with the exception of DUF1620, which is a bigram with PF13360 - a PQQ-like domain (Finn et al., 2014a) found in 186 species that are mostly

bacteria and not reported in fungi. Thus, it is difficult to speculate about the role and function of DUF1620. Additionally, BLASTP searches with these proteins against the NCBI nr database returned *Fusarium* species or the insect pathogen *Metarhizium robertsii* among their top hits.

DUF814, DUF3435, DUF3505 and DUF3669 were found to be present only in FC and FV, whereas the former three DUFs appears in higher copies in FC than in FV (Table E-2 in Appendix E). Moreover, DUF3435 and DUF3669 appear as a single domain in the proteins, while DUF814 and DUF3505 are combined with pfam domains and the latter also forms a homo-brigram in FC. BLASTP searches against the NCBI nr database returned nothing of note, except FCUL_13493 (containing DUF3435) which was unique to *F. culmorum*.

Five DUFs that are only located in FV and not in the other *Fusarium* species used in this study. These are DUF1203, DUF2278, DUF4238, DUF4341, DUF4646 (Table 6-4). Each of these DUFs represents different FV protein (see Table 6-4). However, according to the PFAM version 28.0 (Finn et al., 2014a), DUF4341 has been removed/ or merged into PF04143.

Furthermore, DUF1203, DUF2278 and DUF4238 are domains that are found mainly in Bacteria kingdom and no hit for fungi were detected in the PFAM version 28.0. On the contrary, DUF4646 was found to be a fungal-specific domain and per PFAM version 28.0 is found in the *F. graminearum* proteome. This finding reflects improvement to the FG genome assembly made by Rothamsted Research colleagues (King et al., 2015). Thus, the list of DUFs only present in FV shrinks to: DUF1203, DUF2278 and DUF4238. The latter appears as a hetero-bigram with a PF00932 domain, while the others are the only domain in their respective proteins. The pfam annotation for PF00932 revealed that it is well represented in Bacteria and some Eukaryotes, though not fungi, and BLASTP returned various plant pathogens.

There were two FC-specific DUFs (Table 6-4), both are single-domains in their respective proteins, and pfam states that DUF3106 is predominantly present in bacteria but absent from fungi. Five DUFs were absent from FG3 but present in the others, which improvements in assembly and gene calls for FGGRES and the other two species.

Table 6-4 DUFs identified in only one species.

Species	DUF Id	Pfam Id	Gene	Other domains*	NCBI blastp - the best hits	L [%]	Id[%]
***F.graminearum* (FG3 MIPS gene call)**	DUF1330	PF07045	FGSG_11455	No	*F. graminearum* PH-1 (HP FGSG_11455)	100	100
***F.graminearum* (FGRRE gene call)**			FGRRES_11455		*S. chartarum* IBT 40288 (HP S40288_10353)	85	81
					F. oxysporum f. sp. *pisi* HDV247 (HP FOVG_17228)	85	81
F. culmorum	DUF3106	PF11304	FCUL_09034	No	*F. pseudograminearum* CS3096 (HP FPSE_00972)	100	91
	DUF3723	PF12520	FCUL_09951	No	*F. oxysporum* f. sp. *conglutinans race* 2 54008 (HP FOPG_15645)	99	48
					F. oxysporum Fo5176 (HP FOXB_16449)	20	60
F. venenatum	DUF1203	PF06718	FV_04628	No	*F. avenaceum* (HP FAVG1_10859)	99	78
	DUF2278	PF10042	FV_13774	DUF2278|PF00932	*F. verticillioides* 7600 (HP FVEG_03407)	100	86
					F. oxysporum f. sp. *cubense race* 4 (Putative protein yukJ)	99	78
					F. oxysporum f. sp. *lycopersici* MN25 (HP FOWG_00001)	99	78
	DUF4238	PF14022	FV_08173		*F. oxysporum* f. sp. *raphani* 54005 (HP FOQG_13556)	100	89
					F. oxysporum f. sp. *pisi* HDV247 (HP FOVG_14684	100	88
					F. oxysporum FOSC 3-a (HP FOYG_10460)	100	88
	DUF4341**	PF14241	FV_10781	PF04143|DUF4341|PF04143	*F. graminearum* PH-1 (HP FGSG_11292)	99	91
					F. pseudograminearum CS3096 (HP FPSE_08411)	99	90
	DUF4646	PF15496	FV_03749		*F. pseudograminearum* CS3096 (HP FPSE_12375)	100	78
					F. graminearum PH-1 (HP FGSG_11292)	98	77

*Includes DUF tested. **DUF4341 has been removed/ or merged into PF04143 in PFAM version 28 (Finn et al., 2014a).

Of the fungi-specific DUFs identified in the previous chapter (Table 5-8), 35 were present in all *Fusarium* proteomes (Figure 6-3) and are almost evenly distributed throughout these proteomes (Figure 6-4, Table 6-5). However, three DUFs were overrepresented in FV proteome: DUF3433, DUF3425 and DUF3176.

As mentioned earlier in this chapter, both DUF3433 and DUF3425 were also found to be more abundant within pathogenic fungi and fungal plant pathogens (see Figures 5-11 and 5-12), as well as DUF3176 were among those DUFs representing in higher number pathogenic fungi (see Figures 5-11 and 5-12). Although these three DUFs have been shown to be more abundant, the chi-square test performed in Chapter 5 did prove that these DUFs appear independently within different lifestyles of fungi.

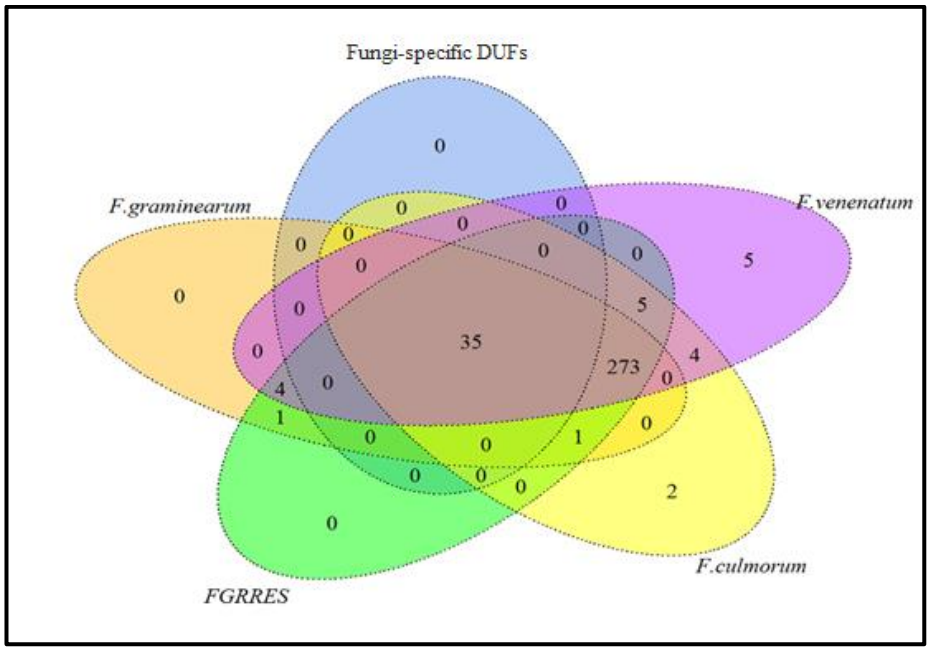


Figure 6-3 Distribution of 35 fungi-specific DUF within four *Fusarium* proteomes.
Where FGRRES set is a Rothamsted Research *F. graminearum* genome assembly, whereas *F. graminearum* set indicates FG3 MIPS genome assembly. Fungi-specific DUFs set groups 35 DUFs identified as fungi-specific DUFs in Chapter 5.

Furthermore, as FV share all 35 fungi-specific DUFs within plant pathogenic fungi, Table 5-10 from Chapter 5 can be redrawn including FV as second non-infecting fungus and both *F. graminearum* (replacing FG anamorph: *Gibberella zeae*) and *F. culmorum* as plant pathogenic fungi (Table 6-6).

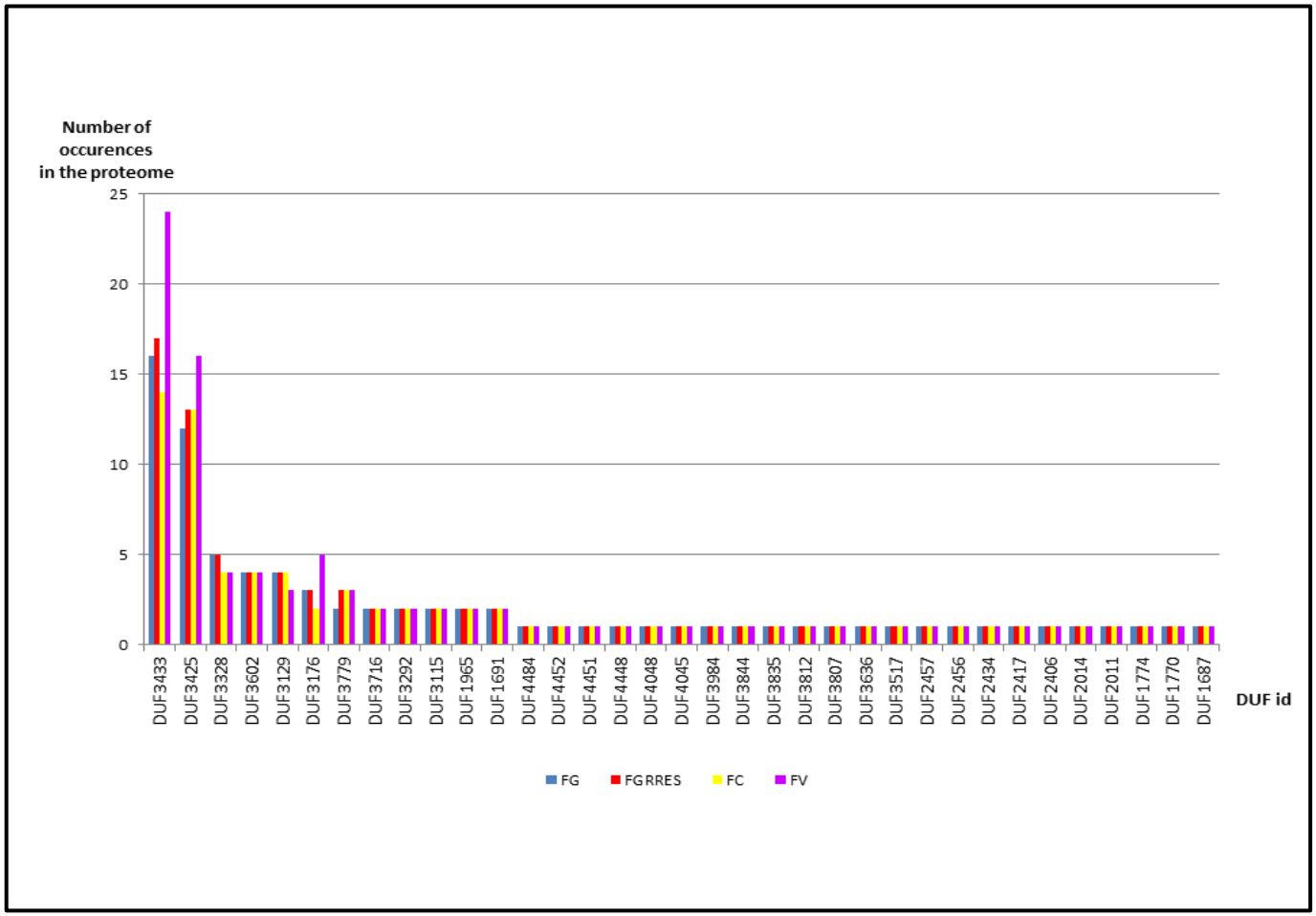


Figure 6-4 Distribution of DUFs specific to fungi across tested *Fusarium* proteomes.
Where FG – *F. graminearum* FG3 MIPS gene call, FGRRES - Rothamsted Research *F. graminearum* gene call, FC – *F. culmorum* and FV – *F. venenatum.*

Table 6-5 Fungi-specific DUFs across pathogenic and non-pathogenic proteomes.

No	DUF Id	FG	FGRRES	FC	FV
1	**DUF3433**	16	17	14	24
2	**DUF3425**	12	13	13	16
3	**DUF3328**	5	5	4	4
4	**DUF3602**	4	4	4	4
5	**DUF3129**	4	4	4	3
6	**DUF3176**	3	3	2	5
7	**DUF3779**	2	3	3	3
8	**DUF3716**	2	2	2	2
9	**DUF3292**	2	2	2	2
10	**DUF3115**	2	2	2	2
11	**DUF1965**	2	2	2	2
12	**DUF1691**	2	2	2	2
13	**DUF4484**	1	1	1	1
14	**DUF4452**	1	1	1	1
15	**DUF4451**	1	1	1	1
16	**DUF4448**	1	1	1	1
17	**DUF4048**	1	1	1	1
18	**DUF4045**	1	1	1	1
19	**DUF3984**	1	1	1	1
20	**DUF3844**	1	1	1	1
21	**DUF3835**	1	1	1	1
22	**DUF3812**	1	1	1	1
23	**DUF3807**	1	1	1	1
24	**DUF3636**	1	1	1	1
25	**DUF3517**	1	1	1	1
26	**DUF2457**	1	1	1	1
27	**DUF2456**	1	1	1	1
28	**DUF2434**	1	1	1	1
29	**DUF2417**	1	1	1	1
30	**DUF2406**	1	1	1	1
31	**DUF2014**	1	1	1	1
32	**DUF2011**	1	1	1	1
33	**DUF1774**	1	1	1	1
34	**DUF1770**	1	1	1	1
35	**DUF1687**	1	1	1	1

Where FG – *F. graminearum* FG3 MIPS gene call, FGRRES – *F. graminearum* Rothamsted Research gene call, FC – *F. culmorum* and FV – *F. venenatum*. Highlighted in grey DUFs domains with discrepancy in a number of occurrences within pathogenic and non-pathogenic fungi proteomes.

Table 6-6 Fungi with the whole repertoire of DUFs specific to fungal species – improved.

No	Plant pathogens possessing all 35 DUFs	Fungi infecting fungi with all 35 DUFs	Symbionts of plants roots and endophyte	Not infecting Fungi with all 35 DUFs
1	*Colletotrichum gloeosporioides*	*Hypocrea atroviridis*	*Pestalotiopsis fici*	*Podospora anserina*
2	*Colletotrichum graminicola*	*Hypocrea virens*		*Fusarium venenatum*
3	*Fusarium culmorum*			
4	*Fusarium graminearum*			
5	*Fusarium oxysporum*			
6	*Fusarium pseudograminearum*			
7	*Gibberella fujikuroi*			
8	*Gibberella moniliformis*			
9	*Magnaporthe oryzae*			
10	*Nectria haematococca*			

Finally, the occurrence of 35 fungi-specific DUFs in all sets of proteomes in this study indicates that we are working with high quality of genomic and proteomic data and also helps in some respect to validate that we are dealing with the good and deep representation of the genomes, especially FGRRES and FV genomes.

6.4 Discussion

In this chapter, pfam-domain repertoires have been compared and predicted for four *Fusarium* proteomes. The main emphasis was placed on their content, abundance, and diversity between plant fungal pathogen, namely *F. graminearum* and *F. culmorum* (FC) and non-pathogenic fungi represented by *F. venenatum (*FV). The analysis in this chapter employs the methodology used in Chapter 5 and extends it further towards other species of the same genus. In addition, as a new gene call for *F. graminearum* was available at the time of writing this chapter, the DUF repertoire of the previous gene call (FG3 MIPS) was compared to that of the newly available FGRRES proteome. Thus, four genomes were compared.

Firstly, the total overview of pfam domain and DUFs among the genomes was undertaken. Overall, pfam domain content (including DUFs) is similar for four proteomes in this study, ranging from the lowest in FC and the highest for FV. Obviously, the content of pfam domains in the new FGRRES proteome is higher than in the previous gene call of FG. This, of course reflects the improvement made into the FG genome assembly and the annotation (King et al., 2015).

Moreover, the lowest number of domains in the FC proteome reflects the FC fragmented genome assembly containing 3M unspecified bases.

DUFs, however, represent only a small percentage (~4%) of the total pfam domains for all four proteomes. The highest number of DUFs (538) was observed within the FV proteome, whereas the lowest number of DUFs was detected in both FG and FGRRES proteomes (504 and 515 respectively). The same trend is also observed while comparing the number of proteins with only one pfam domain that is DUF. The difference in the DUF content between FG /FGRRES and FV might suggest that the latter, as a newly assembled and annotated genome, which is not yet publicly available, has more proteins that were not experimentally tested for their function, which in turn reflects a higher number of DUFs within its proteome. On the other hand, it is possible that the higher content of DUFs within non-pathogenic fungi can be related to its saprophytic lifestyle. Thus, as suggested in the previous chapter, proteins consisting of only one domain that is DUF may be essential for life and not necessary for pathogenicity.

Furthermore, the total percentage of pfam domains that are DUFs is also slightly higher for FC, which might reflect the novelty or the fragmentation of the genome (Table 6-1). However, the number of proteins with only one DUF domain is much lower (308) in FC compared to other proteomes. This might suggest that DUFs in FC are likely to form bigrams with other pfam domains and the function of some might be a result of their interaction with other domains (Vogel et al., 2004a). Nevertheless, further analysis, including a comparative study of bigrams across several species, will need to be performed to validate this hypothesis.

One DUF, namely DUF1330, was only found in *F. graminearum* proteome. This DUF also exists as a one-domain-only protein (FGSG_11455 /FGRRES_11455). Furthermore, BLASTP only retrieved the *F. graminearum* protein FGSG_11455, indicating that this protein might be FG specific. Further examination of DUF1330 with the help of PFAM version 28, disclosed that this domain is common across bacterial species and is only found in the fungal species *F. graminearum* and *F. oxysporum*. However, FGSG_11455 has neither been experimentally tested nor predicted as a candidate gene for pathogenicity in the study by Lysenko et al. (2013).

Moreover, DUF3723 and DUF3106, which are only present in FC and are part of single domain proteins FCUL_09951 and FCUL_09034 respectively, are suggested to be FC specific. Additionally, one of the FC proteins, namely FCUL_13493 protein, which shares a DUF3435 domain with FV, has been suggested to be an FC specific as nothing similar was found in the NCBI nr database.

Although some discrepancy in the DUF content across the four proteomes was detected, 308 DUFs were found to be in common. Out of these 308 DUFs, 35 fungi-specific DUFs (identified in Chapter 5) were found. Furthermore, these 35 DUFs are almost evenly distributed throughout the four proteomes in this study. An exception here are three DUFs, namely DUF3433, DUF3425 and DUF3176 which are slightly more overrepresented in FV. Although these three DUFs were also found to be more abundant within plant pathogenic fungi (see Figure 5-11 and 5-12 in Chapter 5), the chi-square test performed in Chapter 5 verified that these DUFs appear independently within different fungi lifestyles.

Finally, the majority of DUFs detected in this study are part of genes/ proteins for which similar protein sequences were found in other *Fusaria* proteomes representing plant pathogenic fungi. However, this does not indicate that DUFs are more common among pathogenic species, since many species of the genus *Fusariu*m are fungal pathogens. The findings in this chapter confirm the outcome from Chapter 5 that there are DUFs specific to fungi. Moreover, these DUFs are evenly distributed among different species of the same genus and the occurrence of these DUFs within the proteome is not affected by fungal lifestyle. Thus, their function is not directly connected to the pathogenicity, despite the small evidence (result of chi-square test in Chapter 5) that some DUFs are related to the fungal lifestyle.

Chapter 7

Different types of network analysis to explore various plant pathogen interactions and lifestyles

Both analyses in Chapters 5 and 6 concentrated mainly on functional units of protein – domains and their contribution to a protein functional annotation. The main attention in these chapters was given to DUFs and their possible role in the pathogenic lifestyle of fungi. In addition, the domain co-occurrence network was investigated for prediction of DUF function and to unlock possible functions of proteins built from DUF(s). However, in studies conducted in Chapters 5 and 6, it was not possible to define the specific role of DUFs in pathogenic fungi. Only for small proteins containing one DUF and no other domains, these were shown to be essential proteins (lethal to the organism when protein is silenced or knocked out).

As mentioned in Chapter 2, most domains are not exclusive to only one protein but instead the same domain can occur in a variety of proteins including non-homologues ones. This might suggest that the function of a single domain is reused in several proteins (Buljan and Bateman, 2009). This can easily be explained because new proteins evolve through duplication and recombination of a narrow set of domains. In addition, a small point mutation in the protein domain may have a considerable effect on its function (Chothia et al., 2003).

Most domains in eukaryotes are found in an identical arrangement in several proteins suggesting the combination have descended from a common ancestor (Chothia et al., 2003). However, some domains form a variety of combinations with other domains and are considered as promiscuous. This might be because some domain architectures result from selective forces that facilitate them to stay in a population (Buljan and Bateman, 2009). The study by (Basu et al., 2008) suggests that promiscuous domains are involved in protein-protein interactions and contribute to signaling pathways.

Although homology-based methods are widely used computational tools for functional annotation of proteins and their domains, some domains like DUFs do not share significant sequence

similarity and/ or structure with well characterised domain families to ascertain their molecular function via homology detection tools. On the other hand, the analysis in Chapter 5 and another independent study (Mudgal et al., 2015) reveals that DUF signatures show some conservation across all main kingdoms of life. Moreover, some DUF families showed only conservation among the narrow taxonomical group like fungi (35 fungi-specific DUFs discovered and described in Chapter 5 and 6) or bacteria (Goodacre et al., 2014).

The study by Mudgal et al. (2015) showed that utilisation of different remote similarity detection methods can be used to annotate families with DUFs. The detection methods described in the study assisted structural annotation of DUFs, which could then suggest molecular functions for DUFs. The outcome of the study clarifies to some extent why DUFs in the domain-association network, created and studied in Chapter 5, are defined as non-hub ultra-peripheral nodes that do not share a connection and / or information with the common, hub-like (important) nodes in the network.

Considering the evolution of protein domains in the genome, DUFs are an example of domains whose sequences can be altered or completely changed during the evolution by mutations, deletions, and insertions. This led to new domains with a structure and/or function different to the original domain. Thus, difficulties can be anticipated to predict their function based only on simple sequence similarity methods including the domain-association network analysis.

7.1 Introduction

In the domain co-occurrence network analysis (the domain-association network established in Chapter 5), the function of a multiple domain protein could be predicted by integrating the function of each domain. As it has been demonstrated in Chapter 5, the task of prediction of multiple domain protein function by integrating functions of different domains located within a single protein is not straightforward. This is clearly depicted in Figure 5-14 (Chapter 5), where some nodes (domains) in the network inherited contradictory phenotype annotations, for example ‘Unaffected pathogenicity/Reduced virulence’. This indicates that the same domain is present in FG proteins

experimentally verified to affect the pathogenicity and in proteins which have not been implicated in the pathogenic lifestyle.

To avoid the inconsistency delivered by the domain-association network analysis, in this chapter different types of PPI networks are explored to predict or elucidate unknown functions of proteins in plant pathogenic fungi. There are three components (concepts) to this study.

In the first of them, the pre-existing predicted interactomes for plant fungal pathogens and their hosts are utilised to predict the host-pathogen PPI (HPPPI) networks. Predicting interspecies protein-protein interactions between fungal pathogen and its host are vital for understanding the mechanism by which a pathogen infects the host. Furthermore, predicting the host-pathogen interactions system is essential not only for developing a better treatment but also plays an important role in prevention of the diseases caused by a pathogen. Since experimental techniques remain expensive and time consuming, the computational prediction of HPPPI still continues to be an important activity in proteomics.

There are several computational approaches for prediction of HPPPI: homology-based (or interologs approach) (Lee et al., 2008, Wang et al., 2013), domain interactions (Dyer et al., 2007, Nguyen and Ho, 2008, Zhou et al., 2013), and machine learning (Barman et al., 2014, Jansen et al., 2003).

The second concept of this study involves exploring shared bigrams (an ordered pair of pfam domains) across the same fungal genus, whereas in the third concept shared bigrams and orthologs are explored within the wider fungal taxonomy.

In the second component of the study, the network approach methodology, adopted from the study by Liang et al (2013), was modified and applied to create a Protein Bigrams Overlap Network (PBON) comprising of proteins from three *Fusarium* species, including two plant pathogens, namely *Fusarium graminearum*, *Fusarium culmorum*, and non-pathogenic fungus *Fusarium venenatum.*

Finally, in the third component of this study, the concept of PBON application is further employed in the integration of protein sequences from a wider group of fungal species representing plant

pathogenic fungi *(F. graminearum* and *M. oryzae*), as well as the fungus with saprophytic lifestyle (*N. crassa*). Then, the clustering method SimMod (Bennett et al., 2015) was adopted and used to detect composite modules between three PBONs created for *F. graminearum*, *M. oryzae* and *N. crassa*. In addition mutant phenotype data obtained from PHI-base version 3.8 (Urban et al., 2015b) and the BROAD phenotyping platforms for *N. crassa* were applied to further investigate the biology of the generated clusters. This part of the study primarily concentrates on the clusters forming of proteins from only pathogenic species, namely *F. graminearum* and *M. oryzae* in order to be able to speculate about possible phenotypic outcomes for proteins not tested for a function.

Each of the three concepts studied in this chapter has a particular aim and objective listed in section 7.2.

7.2 Aims and objectives

The aim of the first concept is to predict a HPPPI network for *F. graminearum* and *M. oryzae* and rice as their host.

The aim of the second concept is to identify pathogenic protein clusters and species-specific proteins across the *Fusarium* genus including both pathogenic and non-pathogenic fungi. The intention of this analysis is to discover possible, if any, diversity in the protein repertoire of *Fusarium* fungi with pathogenic and non-pathogenic lifestyles.

The aim of the third concept is to improve the prediction of pathogenic protein clusters and species-specific protein clusters by going to the broader taxonomical space of fungi and including both pathogenic and non-pathogenic species.

7.3 Resources and methods

7.3.1 Prediction of host-pathogen interactions

The analysis was motivated by the study by Mukhtar et al. (2011), where physical interactions of *Arabidopsis thaliana* were mapped to effectors proteins from two pathogens, namely

Hyaloperonospora arabidopsis (Hpa) and *Pseudomonas siringae* (*Psy*). As a result of the study, an experimentally predicted Plant Pathogen Immune Network (PPIN-1) was generated.

For the prediction of HPPPI networks *F. graminearum* and *M. oryzae* were chosen as pathogens and rice was selected as their host. Although *F. graminearum* is the main pathogen of the wheat there is evidence that *F. graminearum* also infects rice (Choi et al., 2015, Gomes et al., 2015, Lee et al., 2009). Therefore, and also due to the lack of completed wheat genome at the time of performing the analysis (June 2014), rice was chosen as a host for both fungal pathogens.

In this section the method proposed to predict HPPPI within two interacting systems, namely *F. graminearum* and rice, as well as *M. oryzae* and rice is described. There are already predicted interactomes for both fungal plant pathogens, and rice available (Zhao et al., 2009, He et al., 2008, Zhu et al., 2011).

7.3.1.1 Test model

The data from the previous study by Mukhtar et al. (2011) were intended to be used in my prediction of HPPPI. They generated an immune interaction network (PPIN-1) between *Arabidopsis thaliana* (host) and *Hyaloperonospora arabidopsis* (pathogen 1) and *Pseudomonas siringae* (pathogen 2) using experimentally validated data from yeast-two-hybrid screens. The PPIN-1 consisted of 1358 interactions between 165 unique proteins from *A. thaliana* and in total 83 effector groups including *H. arabidopsidis* and *P. syringae* effectors genes.

7.3.1.2 Proposed steps of the analysis

The first step was to predict two PPI networks, so called interactomes, separately for *H. arabidopsidis* and *P. syringae* as such did not exist at the time of performing the analysis in June 2014. Here, both ortholog (interologs) and domain-domain approaches were suggested for interactome prediction.

There is already an experimentally verified interactome for *A. thaliana* available (Geisler-Lee et al., 2007). With the three interactomes predicted separately, HPPPI between *A. thaliana* and *P. syringae*, and HPPPI between *A. thaliana* and *H. arabidopsidis* could be predicted using either or

both methods suggested above, namely interologs and / or domain-domain approaches. Once interacting proteins are predicted for *A. thaliana* and *P. syringae*, as well as for *A. thaliana* and *H. arabidopsidis* systems it was proposed to map effectors genes and their interacting proteins from the Mukhtar et al. (2011) study into the predicted host-pathogen interaction systems. The number of correctly predicted interactions between host and each pathogen separately compared to the number of PPI predicted in Mukhtar et al. (2011) study would give some confidence in the interaction prediction methods proposed in this study to predict host-pathogen interaction.

The next step would have been to compare the confidence levels between the interaction prediction methods, namely orthologs and domain-domain approaches. Then, the better method (with the higher confidence level) would be applied to predict host-pathogen interactions between *M. oryzae* and rice, as well as host-pathogen interactions between *F. graminearum* and rice.

7.3.2 Data acquisition for the second and the third concepts of the study

Protein sequences for *Fusarium culmorum* (FC) and *Fusarium venenatum* (FV) were downloaded from Rothamsted Research internal resources. In addition, as a new *Fusarium graminearum* (FGRRES) genome assembly and annotation became available at the time of writing this chapter (King et al., 2015), the proteins of the new improved version of the FG genome, as well as protein sequences for *Magnaporthe oryzae* (MO) and *Neurospora crassa* (NC) were downloaded from Ensembl fungi (Kersey et al., 2014).

The pfam domain repertoires of FGRRES, FC, FV, MO and NC were identified using the same methodology and PFAM version 27.0 as described in Chapter 5. The domain overlapping issues in the above five predicted proteomes were also solved by applying the pipeline program (https://github.com/ejsejda/PhD_thesis-Chapter_5/solving_domains_overlapping.py) employed in Chapter 5 (Figure 5-2) and described in section 5.3.1 of Chapter 5.

7.3.3 Construction of Protein Bigrams Overlap Network (PBON)

In PBON each node represents a single protein. Two nodes are connected by an edge if the corresponding proteins share a mutual ‘ordered bigram’. As per Chapter 5, the domain bigram

definition was adopted from the previous study (Seidl et al., 2011) as two successively located domains in a given protein.

In the current analysis, in order to achieve higher similarity between protein sequences, the order of domains pair in the bigram with respect to N/C-terminus was considered to be important (so called 'ordered bigram'), similar to study by Seidl et al. (2011) (see Figure 2-1). Thus, the bigram AB is not the same as the bigram BA and both are defined as 'hetero-bigrams' regarding the content. Repeated domains were also considered in this analysis. Therefore, neighbouring domains A and A would count as an AA bigram classified as 'homo-bigram'. Figure 7-1 summarises the metrics used in PBON network construction.

7.3.3.1 Construction of a composite PBON for *Fusarium* species proteomes

A composite PBON network was constructed for FGRRES, FC and FV. Firstly, the unique 'ordered bigrams' were identified within each proteome. Then, these bigrams were combined into one big unique 'ordered bigram' set, followed by removal of bigram duplicates arising from merging unique 'ordered bigrams' from three proteomes. For each unique 'ordered bigram' in the composite set, the list of protein sequences with the particular bigram within, from all three species was identified and duplicates were removed. Thus, nodes with self-loops were not included in the composite PBON.

Finally, connections between protein sequences sharing the same bigram were made. The weight on the edge connecting these proteins (see Figure 7-1 III) indicates the number of the same 'ordered bigram' shared by connected proteins. Therefore, the higher the weight on the edges the greater similarity of the connected proteins.

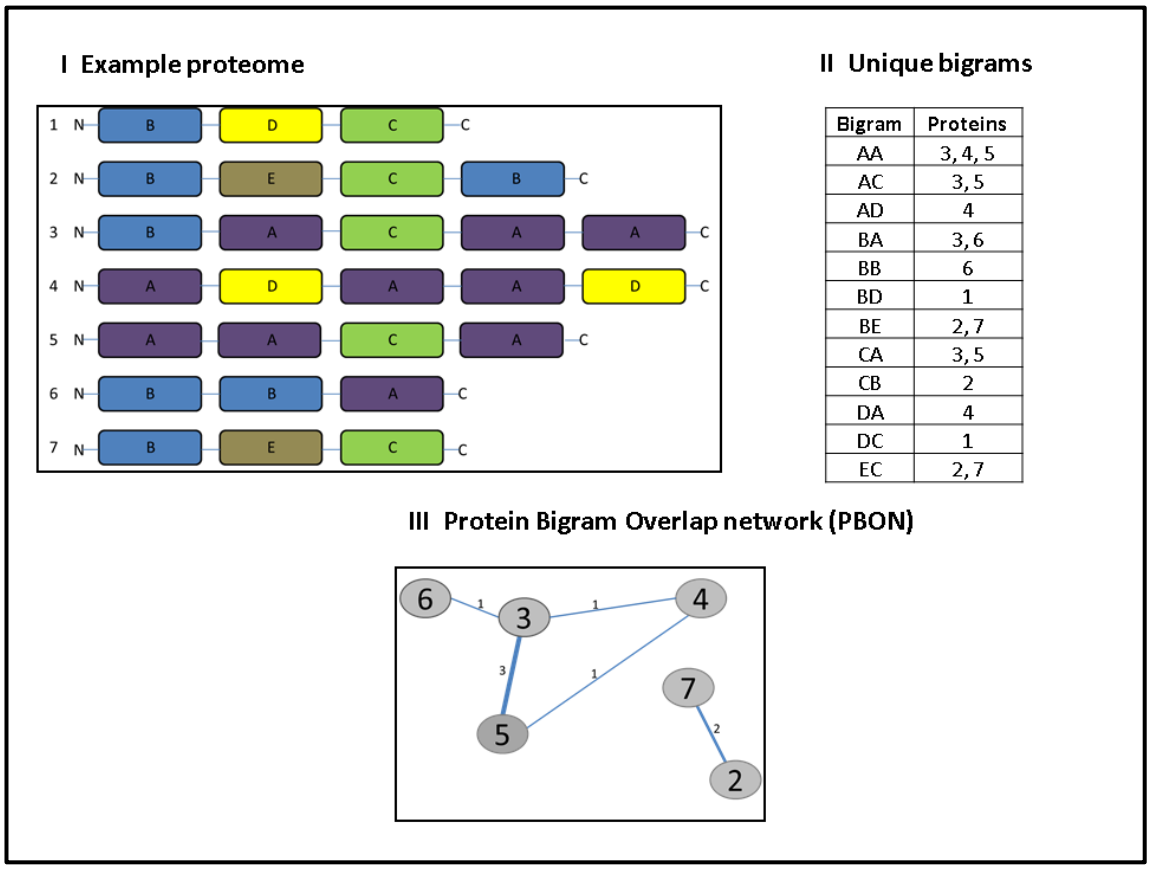


Figure 7-1 Metrics used for construction of Protein Bigrams Overlap Network.
I. An example proteome (seven protein sequences). Each protein is built from pfam domains depicted in different colours and assigned a capital letter. II. Table showing which proteins contain the specified bigram. III. the Protein Bigram Overlap Network (PBON) derived from the example proteome. Nodes in PBON corresponds to protein numbers in an example proteome. The edges of the PBON indicate shared bigram(s) in common between connected proteins. The thickness of the edge indicates the number of shared bigrams. The higher the number of shared bigrams, the thicker the edge connecting proteins.

7.3.3.2 Construction of a composite PBON for fungi from the broader taxonomical space

The concept of PBON application was further intended to be employed in the integration of protein sequences from a wider group of fungal species representing plant pathogenic fungi, as well as the fungi with saprophytic lifestyle or model fungi.

Initially, eight fungal species were chosen for this analysis. These are listed in Table 7-1. However, due to the technical constraints visualising and analysing a combined PBON for the proposed species in Cytoscape, the list of species was reduced to three not closely related ascomycetes species and includes both pathogenic *(F. graminearum* and *M. oryzae*) and non-pathogenic (*N. crassa*).

Table 7-1 List of fungal species intended for further PBON analysis.

No	Species	Lifestyle
1	*Fusarium graminearum* (FGRRES, Rothamsted Research gene call)	plant pathogen (not leaf pathogen)
2	*Fusarium culmorum*	plant pathogen (not leaf pathogen)
3	*Fusarium venenatum*	non-pathogen
4	*Gaeumannomyces graminis* var. *tritic*i R3-111a-1	plant pathogen (infecting roots)
5	*Magnaporthe oryzae* 70-15	plant pathogen (infecting entire plant)
6	*Magnaporthe poae* ATCC 64411	leaves -grass pathogen
7	*Neurospora crassa* OR74A	saprophytic fungus / model fungus
8	*Saccharomyces cerevisiae*	non-pathogen, model fungus

7.3.4 Detection of composite modules in a Combined Protein Bigrams Overlapping Network (CPBON)

Three Protein Bigrams Overlapping Networks (PBONs) were built separately for three ascomycete species, namely *Fusarium graminearum* (FGRRES)*, Magnaporthe oryzae* (MO), and *Neurospora crassa* (NC). Once the PBON was built for each of the species, the orthologs for proteins among these three species were downloaded from BioMart (Smedley et al., 2015) via Ensemble Fungi (Kersey et al., 2014).

For each protein that shares orthologs with either one or both species in this analysis the protein Id was replaced with a node number (n) (see Table 7-2 for details). Therefore, proteins that share orthologs within all three species were assigned individual n numbers in the range from n1 to n3901 in such a way that the n number was the same for three ortholog proteins. If a protein, for example, belongs to FGRRES and share an ortholog protein from only the MO network, then for both proteins FGRRES and MO the same n number was defined from the range of n3902 to n4839.

Table 7-2 Node labeling system for orthologs proteins.

Node id range	Orthologs in species	Node colour
n1 - n3901	FGRRES, MO, NC	yellow
n3902- n4839	FGRRES, MO	cyan
n4840 - n5433	FGRRES, NC	orange
n5434 - n6401	MO, NC	brown

Where FGRRES – *Fusarium graminearum* (Rothamsted Research gene call), MO- *Magnaporthe oryzae* and NC- *Neurospora crassa*.

7.3.4.1 SimMod clustering

SimMod is a mathematical programming model / algorithm for identification of composite modules developed by Dr Laura Bennett (a former fellow PhD student at King's College London (KCL))(Bennett et al., 2015). The algorithm was adopted and applied in this study to ascertain and combine the biological information from PBONs constructed for three fungal species: FGRRES, MO and NC.

In the study by Bennett et al. (2015), the algorithm was used to cluster biological networks build for the same species, namely yeast. In the current study, however, the application was extended further to cluster three biological networks of different species to discover similarities and / or diversities between fungi of different biological lifestyle. Based on the ortholog connections (the same n id) between the three networks, the composite modules were generated. Prior to the running of the algorithm, a specific input file had to be created according to the method described in the study by Bennett et al. (2015).

All three networks were considered to be weighted networks with the weighted degree of nodes in the network. Therefore, for example if a node 'n1' is connected to 'n2' and 'n3' with edge weights of 5 and 3 respectively, then the weighted degree of the node 'n1' is equal to the sum of the weights of all edges from/to this node and in this example is equal to eight.

Once the input file was prepared, it was then given to the fellow PhD student at Kings College London, Jonathan Silva, who performed the run of the SimMod model. The number of maximum 100 modules was chosen based on the several runs of the SimMod method starting with a maximum number of modules m equal to 5, then 10 followed by 20, 50, 75, 100, 150 and 200.

Each time the number of communities generated via SimMod in the first largest connected component of CPBON was compared to the number of communities in the first largest connected component generated by the Louvain clustering method applied to the composite network, created by the connection of the above three networks based on the mutual orthologs prior to the clustering.

With the Louvain clustering method, 14 communities (modules) were distinguished within the first largest connected component, whereas with applying the SimMod clustering with maximum modules number m set to 100, 17 communities were uncovered within the first largest connected component of the CPBON network. The further increase in the number of maximum modules generated m to 150 and then to 200 did not improve the number or the content of the modules generated with m set to 100. In fact, once setting the maximum modules number m to 200, the number of communities within the largest connected component decreases to 15.

7.3.5 Association of phenotypic outcome to the protein nodes in the CPBON

FGRRES and MO phenotypes were acquired from the PHI-base version 3.8 (Urban et al., 2015b), whereas the phenotypic outcome for NC proteins was assigned based on the information from the BROAD Institute.[36]

In general, six phenotypic outcomes defined by PHI-base were distinguished for FGRRES and MO proteins. These are lethal, affecting pathogenicity (grouping together both reduced virulence and loss pathogenicity phenotypes), not affecting pathogenicity (unaffected pathogenicity), effector gene, increased virulence, and mixed phenotype (where different mutant phenotypes were observed on different hosts or host tissue types).

NC proteins were assigned one phenotypic outcome which combines the following phenotypes: abnormal, reduced or lack of conidiation, ascospore, perithecial, aerial hyphae, and protoperitheicia formation.

[36] http://www.broadinstitute.org/annotation/genome/neurospora/Phenotypes.html

Once the three networks, belonging to the three species: FGRRES, MO and NC, were clustered with the aid of the SimMod method and then connected via shared ortholog proteins, the CPBON was visualised in Cytoscape version 3.2.1.[37] Figure 7-2 illustrates the steps in the construction of the CPBON with the SimMod method.

Nodes from different species were distinguished by different colours as follows: FGRRES nodes highlighted in blue, MO nodes in purple and NC showed as green. Moreover, nodes that share orthologs proteins within other species were decorated with other colours depending on which species orthologs they are (see Table 7-2 for details).

In addition, different phenotypes associated with pathogenic fungi were distinguished by the different shape of the node, whereas the associated phenotype for NC was indicated by red boarder around the node (see legend for Figure 7-6). Also, GO annotation information was linked to nodes via BioMart (Smedley et al., 2015). Next, the attention was placed on the composite modules generated with the help of the SimMod algorithm. Primarily, the pathogenic composite modules were further investigated. Pathogenic modules are defined in this study as the modules (clusters) comprising of proteins from FGRRS and MO species.

7.3.6 Data analysis

Topological properties of the networks in the section 7.4.2 of this chapter, as well as the properties of each vertex, namely node degree and clustering coefficient were calculated with NetworkX python package (as per Chapter 5), whereas statistical test comparing nodes properties were carried out using R software version 3.03.

[37] http://www.cytoscape.org/

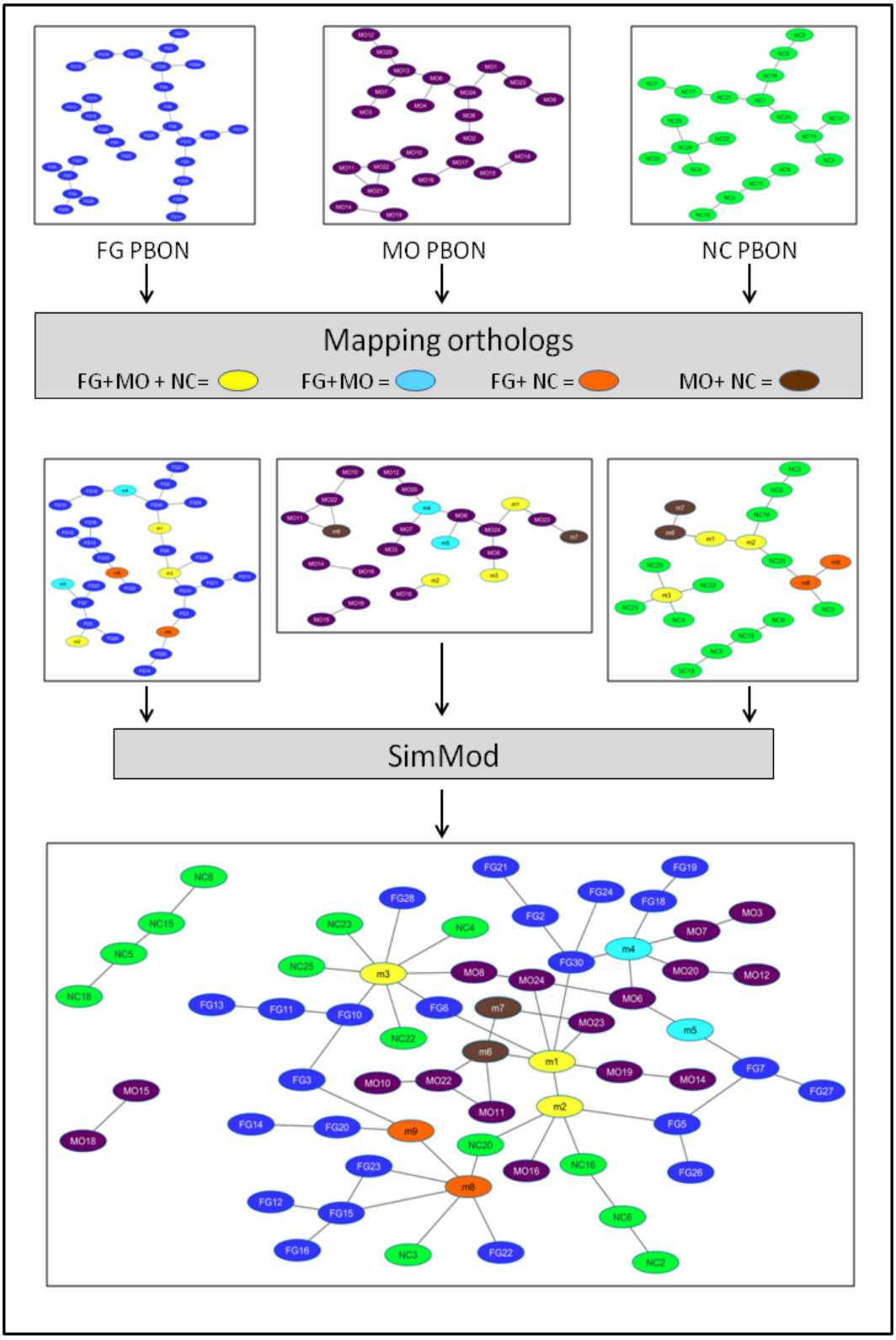


Figure 7-2 Steps in building the Combined Protein Bigrams Overlapping Network and clustering via the SimMod algorithm.

FG – *F. graminearum* (FGRRES – Rothamsted Research gene call), *MO – M. oryzae, NC – N. crassa,* PBON – Protein Bigram Overlapping Network. Nodes colours: blue -FG, purple – MO, green – NC, yellow – orthologs in common for FG, MO and NC, cyan - orthologs in common for FG and MO, orange - orthologs in common for FG and NC, brown - orthologs in common for MO and NC.

7.4 Results

This section is divided into three components (sub-sections) representing three different concepts employed to help with prediction or elucidation of unknown function of proteins in plant pathogenic fungi. The key findings from each component of this study are described in order in this section.

7.4.1 Host pathogen protein-protein interactions prediction

Fusarium graminearum and *Magnaporthe oryzae* were chosen as pathogens and rice was selected as their host in order to predict interspecies protein-protein interactions.

7.4.1.1 Initial investigation

In this part of the study an attempt was made to use the PPIN-1 system generated by Mukhtar et al. (2011) to validate the prediction of PPI between the host and a fungal pathogen. Firstly, it was investigated if it would be possible to map effectors genes from the study by Mukhtar et al.(2011) into the predicted interactomes for *H. arabidopsidis* and *P. syringae* to validate both interactomes.

Detailed analysis of the interactions between *A. thaliana* and effectors genes were reported by Mukhtar et al.(2011). The author reported that there were 83 effector groups interacting with 165 *A. thaliana* proteins. In fact, after detailed analysis of the outcome of the Mukhtar et al. (2011) study, it was found that there are in total 82 different groups of effectors which came from: *H. arabidopsidis* (52) and *P. syringae* (30) interacting with 165 unique *A. thaliana* proteins.

Moreover, while exploring the supplementary table 2 published by the authors, two self-interactions between two effectors groups, namely HARXLL445_group (*H. arabidopsidis* effector) and HOPM1_group (*P. syringae* effector), were observed.

7.4.1.2 Looking for orthologs

The next goal in this part of the study was to look for the orthologs of effector genes in the species for which PPI are experimentally verified. Figure 7-3 depicts species with the highest number of experimentally validated PPI (as per IntAct database in July 2014).

Ideally, to predict PPI for the species of interest, one should choose the reference species which is as close as possible in the tree of life to the species of interest.

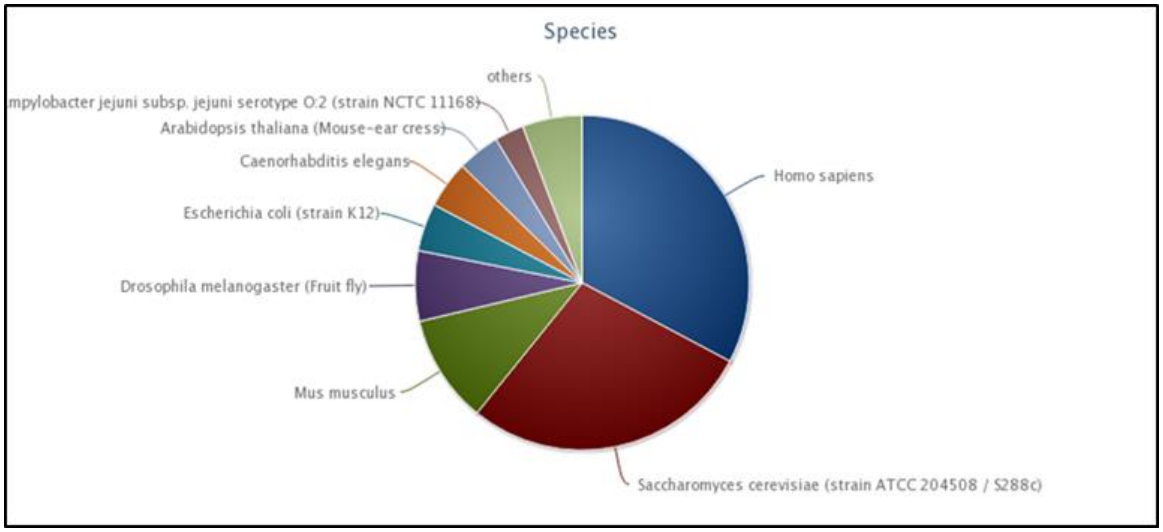


Figure 7-3 The species with the highest number of experimentally verified protein-protein interactions.
Figure copied from IntAct database in July 2014.

The Inparanoid 8 tool and database is the standard resource for finding orthologs. Figures 7-4 shows the overlap between *H. arabidopsidis* and the three largest interactomes. Only two effector genes were found, but these were not present in the PPIN-1 system (Mukhtar et al., 2011). No orthologs were found for *P. syringae*.

Nevertheless, if needed, BLAST search could be performed against well-known species. However, while searching for pfam domains in effector genes listed in the study by (Mukhtar et al., 2011), only 13 out of the total of 30 effector groups for *P. syringae* had a one pfam domain detected. Furthermore, when assessing pfam domains in *H. arabidopsidis* effector groups, only two out of 52 effector groups had one pfam domain detected.

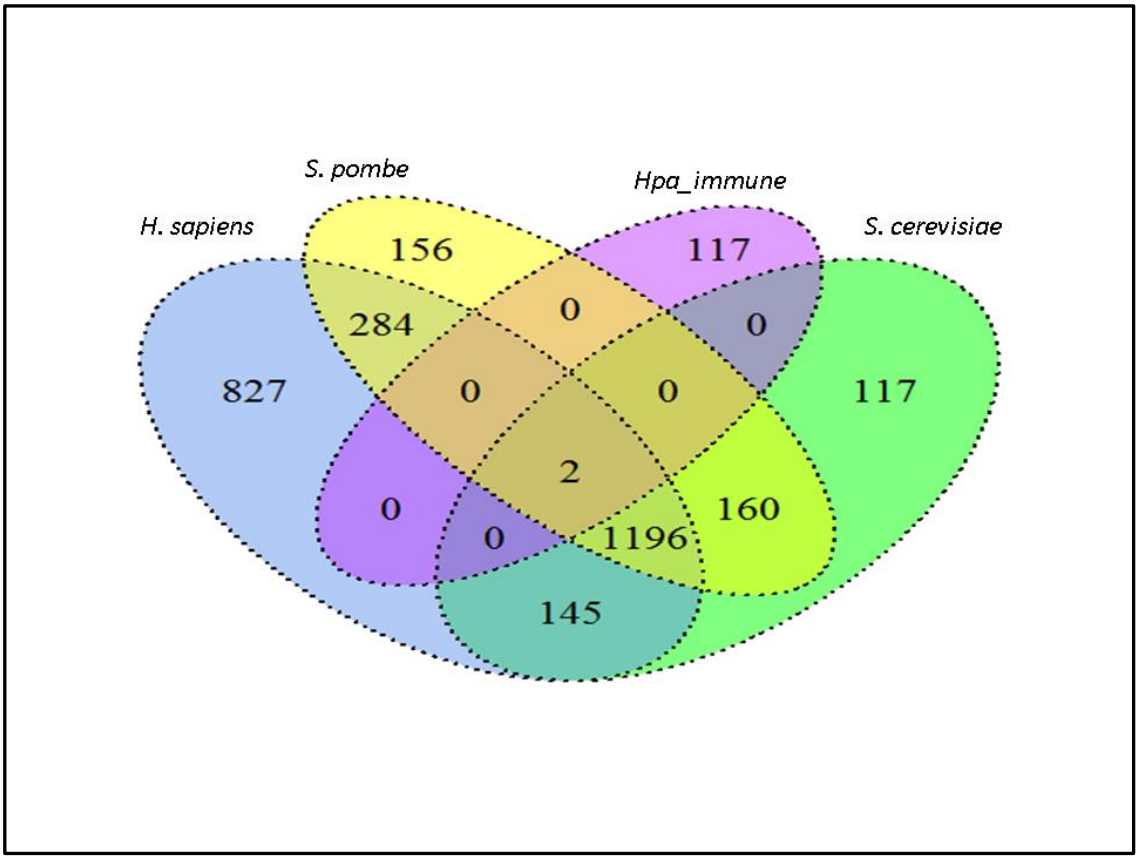


Figure 7-4 Orthologs in common between effector genes in *Hyaloperonospora arabidopsis* (Hpa_immune) and well-studied species.

The initial investigation in this part of the study revealed that effector genes are mostly species-specific. This was evident while searching for orthologs of the oomycete, namely *H. arabidopsidis*. This outcome might also result from the fact that oomycetes are taxonomically far away from fungi and bacteria in the tree of life.

Moreover, it must also be emphasised that while searching for the orthologs of *H. arabidopsidis* against the whole InParanoid 8 database, the majority of hits observed were only for *H. arabidopsidis* proteins and some for other oomycetes such as *Phytophtoria* species. Furthermore, independent investigation of effectors genes in different species (PHI-base version 3.6 analysis) revealed that most of them do not have a pfam domain signature.

7.4.2 *Fusarium* Protein Bigrams Overlap Network (PBON) analysis

Following the unsuccessful approach to predict and validate the HPPPI, it was then considered interesting to explore interactions between proteins of fungi representing both pathogenic and non-pathogenic fungal species. In this part of the study the network approach was employed to investigate the similarities and differences among the protein repertoire of *Fusarium* fungi with pathogenic and non-pathogenic lifestyles, namely *F. graminearum* (FGRRES), *F. culmorum* (FC) and *F.venenatum* (FV).

7.4.2.1 General network characteristics

In total 7,360 proteins across three *Fusarium* species (FGRRES, FC and FV) were employed in the building of the PBON (see Figure 7-5) with 236,561 connections among these proteins. Overall, the participation of proteins from each species in the network was almost uniform ranging from 34% for FGRRES (2,493 nodes) and FV (2,515 nodes) proteins to 32% for FC (2,352 nodes) proteins (see Table 7-3).

Moreover, in all three species, most qualified proteins for contribution in PBON were employed in PBON building. A qualified protein is defined as a protein with at least two domains and therefore can participate in the construction of the PBON. As listed in Table 7-3, almost all FGRRES qualified proteins were included in the network (99.36%), accounting for only 16 FGRRES proteins that were not part of the PBON. A similar trend was observed for FV protein, where 98% of the total qualified proteins were incorporated in the PBON construction, leaving only 44 FV proteins not being part of the network. A slight discrepancy was observed when analysing FC proteins within the network, where 90% of FC qualified proteins contribute to the network formation. This left 260 FC qualified proteins that were absent from the PBON.

Table 7-3 Contribution of species proteins into the PBON construction.

Species	A	B	C	D [%]	E [%]
***F. graminearum* (FGRRES)**	14,164	2509	2493	99.36	33.87
F. culmorum	13,929	2612	2352	90.05	31.96
F. venenatum	13,946	2559	2515	98.28	34.17

Where A – is a size of species proteome, B – number of proteins in the proteome with at least two pfam domains (number of qualified proteins for PBON construction), C – number of proteins within PBON, D = C / B * 100 – the percentage of qualified proteins used in PBON construction, E = C / the total nodes number *100 – the percentage of species proteins in the PBON, where total nodes number in PBON accounts for 7,360.

Identified in the PBON were 931 connected components (CCs). The first and the largest connected component (CC) consisted of 1,377 nodes and most the connections (edges) in the whole network lay within this largest CC (161,886). More than half (546) of the total CCs consisted of three nodes, whilst 90 of the CCs consist of only two nodes. The main metrics of the PBON are listed in Table 7-4.

Table 7-4 Main metrics of the PBON.

Main metrics of PBON	
Number of all nodes in PBON	7360
Number of edges in PBON	236561
Number of connected components	931
Highly connected node in the PBON	FCUL_09968 (node degree = 601)
Highly connected FGRRES node in the PBON	FGRRES_16446_M (node degree = 561)
Highly connected FV node in the PBON	FV_05448 (node degree = 546)
Average clustering coefficient	0.9625
Average node degree	64
Number of FGRRES nodes in PBON	2493
Number of FC nodes in PBON	2352
Number of FV nodes in PBON	2515
Number of nodes in 1st largest CC of the PBON	1377
Number of edges in the 1st largest CC of the PBON	161,886

Where CC – connected component, PBON – Protein Bigrams Overlap Network

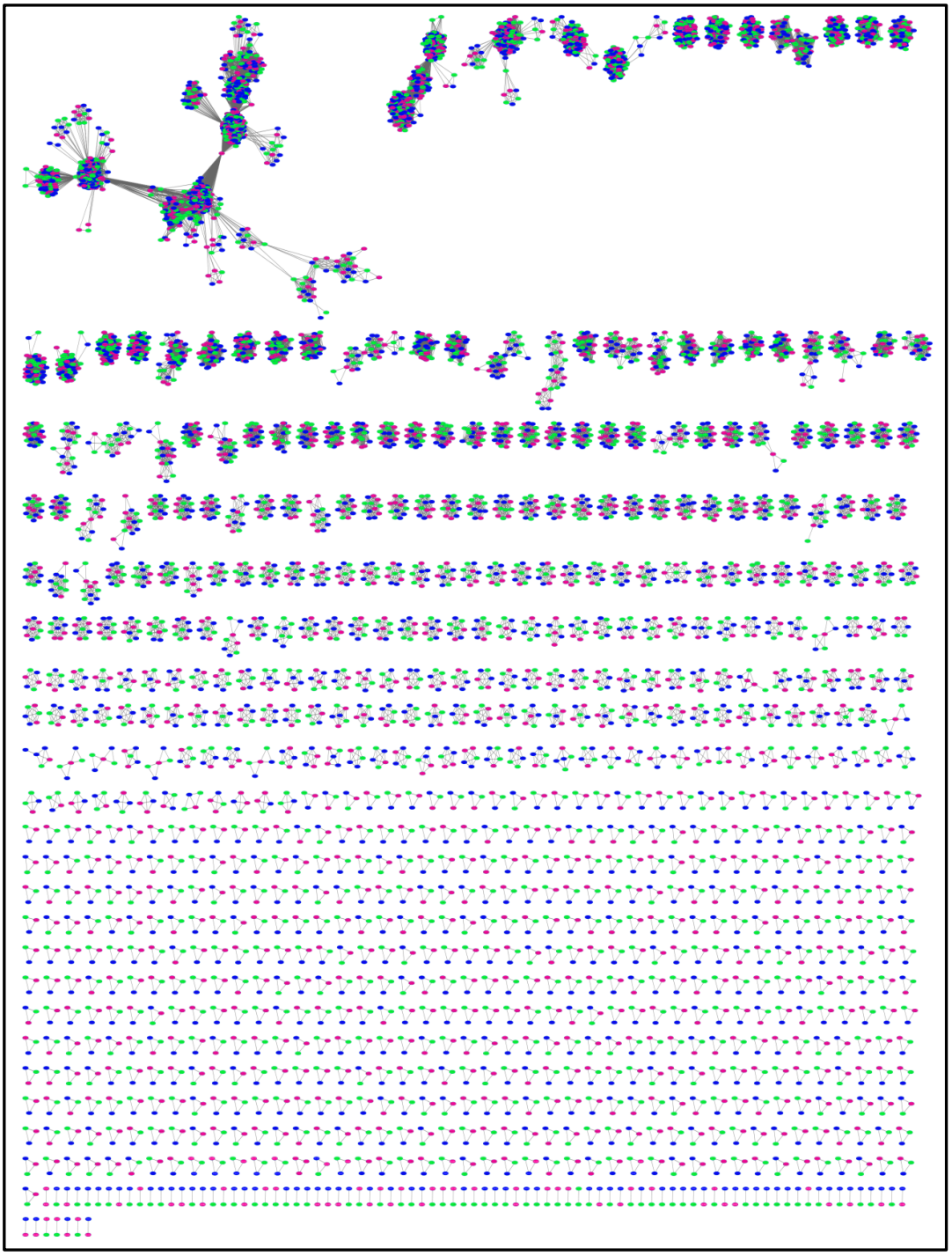


Figure 7-5 Protein Bigrams Overlap Network (PBON) for *Fusarium* species.
Where the edges of the PBON indicate shared bigram(s) in common between connected proteins, the thickness of the edge indicates the number of shared bigrams. The higher the number of shared bigrams, the thicker the edge connecting proteins.

7.4.2.2 Topological properties of the nodes in PBON

Each node in the PBON is associated with one of three *Fusarium* species including two plant pathogen fungi (FGRRES and FC) and one non-pathogenic fungus (FC). Here, it was interesting to examine if a different biological lifestyle of the *Fusarium* species has an influence on the distribution of the node parameters within the network. Thus, two main topological properties of the PBON were investigated in this section. These are node degree distribution and the clustering coefficient.

In order to compare the distribution of these network properties among the three *Fusarium* species, two non-parametric statistical tests were performed. Firstly, the Wilcoxon-Mann-Whitney rank sum test (Wilcoxon Rank Sum Test with continuity correction in R package) was conducted to compare averages of three independent tested groups of nodes. The test revealed that there were no significant differences between nodes in the tested groups (p-value > 0.05 for all tested groups of nodes) for both tested parameters (Table 7-5 A and Table 7-6 A).

Furthermore, a second non-parametric statistical test, the Kolmogorov-Smirnov (KS) test was carried out to re-explore the initial finding. As mentioned already in the section 5.4.5.2 of Chapter 5, KS test is much stronger test than Wilcoxon-Mann-Whitney rank sum test in terms of comparing two independent distributions.

The KS test confirmed the earlier finding of no significant difference (p-value > 0.05) between the distribution of node sets representing each species for both tested parameters (Table 7-4 B and Table 7-5 B).

Table 7-5 Node degree distribution comparison.

A. Wilcoxon-Mann-Whitney rank sum test (Wilcoxon Rank Sum Test with continuity correction in R package)

Node degree							
Nodes		Sample size		Median		W	p-value
1	2	1	2	1	2		
FGRRES	FV	2493	2515	11	11	3106700	0.5790
FGRRES	FC	2493	2352	11	11	2948400	0.7313
FC	FV	2352	2515	11	11	2914400	0.3742

B. Kolmogorov-Smirnov distributions comparison test

Node degree							
Nodes		Sample size		Median		D	p-value
1	2	1	2	1	2		
FGRRES	FV	2493	2515	11	11	0.009785	0.9998
FGRRES	FC	2493	2352	11	11	0.008702	1.0000
FC	FV	2352	2515	11	11	0.016667	0.8883

1 – first data set, 2 – second data set, FGRRES – *F. graminearum* (FGRRES gene call), FC – *F. culmorum*, FV – *F. venenatum*

Table 7-6 Node clustering coefficient distribution comparison.

A. Wilcoxon-Mann-Whitney rank sum test (Wilcoxon Rank Sum Test with continuity correction in R package)

Clustering coefficient							
Nodes		Sample size		Median		W	p-value
1	2	1	2	1	2		
FGRRES	FV	2493	2515	1	1	3114500	0.4579
FGRRES	FC	2493	2352	1	1	2942900	0.6788
FC	FV	2352	2515	1	1	2927200	0.2534

B. Kolmogorov-Smirnov distributions comparison test

Clustering coefficient							
Nodes		Sample size		Median		D	p-value
1	2	1	2	1	2		
FGRRES	FV	2493	2515	1	1	0.0069	1.0000
FGRRES	FC	2493	2352	1	1	0.0061	1.0000
FC	FV	2352	2515	1	1	0.0107	0.9991

1 – first data set, 2 – second data set, FGRRES – *F. graminearum* (FGRRES gene call), FC – *F. culmorum*, FV – *F. venenatum*

Thus, both non-parametric statistical tests revealed that the fungal biological lifestyle of the *Fusarium* species does not necessarily influence the distribution of the node parameters within the network. This is probably because the nodes/ proteins in PBON come from closely related species with high sequences similarity (enormous number of connections in the network)

Therefore, instead of concentrating on the large connected components of the combined network, where proteins (nodes) from different fungal species are evenly distributed, the next step in this analysis was to concentrate on the smallest connected components to find proteins that might be only present in either pathogenic fungi or non-pathogenic fungi.

7.4.2.3 Exploration of similarities within the smaller connected components of PBON

As it was illustrated in Figure 7-5, proteins from the three *Fusarium* species, namely *F. graminearum* (FGRRES), *F. culmorum* (FC), and *F. venenatum* (FV) are evenly distributed within the PBON. In this section, the main attention was focused on the pair of the nodes that are either from the same species or species of the same biological lifestyle. Furthermore, further investigation of the proteins that are not part of the PBON was performed to identify possible orphans within each species.

Out of 90 two-node components (see section 7.4.2.1), 27 pairs were found to be pathogenic fungi-specific and included 25 pairs with both an FGRRES and an FC protein (Table 7-7) and two FC pairs only (Table 7-8). Homologs pairs of FGRRES were not identified. One three-node component with three FC proteins was found (Table 7-9). Each of these triple FC proteins is composed of the DUF3505 homo-bigram.

Furthermore, DUF3505 has been found in two FV proteins, namely in FV_11060 as one copy with the neighbourhood of pfam bigram PF00270|PF00271 and in FV_07932 as a single domain. However, DUF3505 was not detected within the FGRRES proteome, only the completed *Fusarium* proteome (King et al., 2015). Therefore, FC genes: FCUL_13218, FCUL_13198 and FCUL_13213 were not investigated further as ones affecting the pathogenicity of FC. On the other hand, it is formally possible that these three proteins could co-ordinate an FC-specific part of the plant-infection process.

Table 7-7 Two-node connected components with *F. graminearum* and *F. culmorum* proteins.
The table continues on the next page

No	FGRRES protein	FGRRES transcript Id	Protein annotation	Phenotype	FC Id	No	Bigram(s) list
1	CEF85571	FGRRES_06945	related to PAN1 - actin-cytoskeleton assembly protein	Not tested	FCUL_10731	2	PF12763|DUF1720 DUF1720|PF12763
2	CEF83637	FGRRES_17202	hypothetical protein	Not tested	FCUL_12413	2	PF08914|PF01388 PF01388|PF11626
3	CEF76015	FGRRES_17390	hypothetical protein	Not tested	FCUL_03897	2	PF03535|PF12256 PF12256|PF12255
4	CEF72167	FGRRES_00272	conserved hypothetical protein	Not tested	FCUL_00328	1	PF04434|PF12861
5	CEF72844	FGRRES_15789	related to neurofilament triplet H1 protein	Not tested	FCUL_00952	1	PF15624|PF11699
6	CEF73068	FGRRES_01010	related to cold sensitive U2 snRNA suppressor	Not tested	FCUL_01173	1	PF14259|PF13893
7	CEF73778	FGRRES_01614	related to dna mismatch repair homologue (hpms2)	Not tested	FCUL_01830	1	PF13589|PF08676
8	CEF75091	FGRRES_12194	related to GTP-binding protein 2	Not tested	FCUL_03034	1	PF00009|PF03143
9	CEF75316	FGRRES_09994_M	related to nuclear pore protein NSP1 (FGSG_09994)	Not tested	FCUL_03245	1	PF13634|PF00638
10	CEF75412	FGRRES_17311	hypothetical protein	Not tested	FCUL_03338	1	PF13041|PF13041
11	CEF78762	FGRRES_12361_16189_M	conserved hypothetical protein hypothetical protein	Not tested	FCUL_06370	1	PF00651|PF12770
12	CEF79167	FGRRES_12291_2_M	conserved hypothetical protein	Not tested	FCUL_06748	1	PF04828|PF07719
13	CEF79589	FGRRES_04557_M	putative protein [EST hit]	Not tested	FCUL_07131	1	PF00385|PF02178
14	CEF83216	FGRRES_07854	probable YCS4 - subunit of condensin protein complex	Not tested	FCUL_11781	1	PF12922|PF12717
15	CEF83367	FGRRES_09389	related to ADA regulatory protein of adaptive response	Not tested	FCUL_12395	1	PF02805|PF12833
16	CEF83686	FGRRES_09139_40_17163_M	hypothetical protein hypothetical protein	Not tested	FCUL_12726	1	PF01425|PF13527
17	CEF85538	FGRRES_17147	related to tRNA isopentenyltransferase	Not tested	FCUL_12855	1	PF01715|PF12874
18	CEF85961	FGRRES_17581	hypothetical protein	Not tested	FCUL_09832	1	PF00856|PF01753

No	FGRRES protein	FGRRES transcript Id	Protein annotation	Phenotype	FC Id	No	Bigram(s) list
19	CEF86111	FGRRES_06369	conserved hypothetical protein	Not tested	FCUL_09164	1	PF12044|PF01419
20	CEF86218	FGRRES_16596	related to N-acetylglucosaminyl-phosphatidylinositol biosynthetic protein gpi1	Not tested	FCUL_08807	1	PF05254|PF05024
21	**CEF87112**	**FGRRES_06427**	**related to YBR267w**	**RedVir**	**FCUL_09229**	**1**	**PF12756|PF12756**
22	CEF87429	FGRRES_04708	related to CAR2 - ornithine aminotransferase	Not tested	FCUL_07320	1	PF01636|PF00202
23	CEF87749	FGRRES_16593	related to Chromo domain protein Alp13	Not tested	FCUL_08789	1	PF11717|PF05712
24	CEF88289	FGRRES_16573	probable methionyl-tRNA synthetase, mitochondrial	Not tested	FCUL_08654	1	PF09334|PF13302
25	CEF88470	FGRRES_05208	related to nonmuscle myosin-II heavy chain	Not tested	FCUL_07878	1	PF15456|PF02524

Where FGRRES – *F. graminearum* (FGRRES gene call), FC – *F. culmorum,* RedVir – reduced virulence.

Table 7-8 *F. culmorum* two-node connected components.

No	FC protein id	FC protein id	Bigrams No	Bigram(s) list
1	FCUL_05545	FCUL_06148	1	PF12697|PF04082
2	FCUL_02720	FCUL_02959	1	PF07690|PF00067

Where FC – *F. culmorum*

Table 7-9 *F. culmorum* three-nodes connected component.

No	FC	FC	FC	Bigrams No	Bigram(s) list
1	FCUL_13218	FCUL_13198	FCUL_13213	1	DUF3505|DUF3505

Where FC – *F. culmorum*

Moreover, it is not surprising to find paralogous FC genes clustered physically together in the genome, because gene duplication is more common in FC than FG. As expected, no cluster with only FGRRS proteins was found because the repeat-induced point mutation (RIP) mechanism is known to be very active in *F. graminearum* (Cuomo et al., 2007) . RIP is a genome-wide defence system against repeated DNA sequences in the genome, that detects and causes mutation of repetitive sequences and often results in epigenetic silencing of the mutated sequences via DNA methylation (Galagan and Selker, 2004).

Interestingly, one of FGRRES gene/ protein, namely FGRRES_06427 has been experimentally verified to be responsible for the virulence of FG (PHI-base version 3.8). FGRRES_06427 only shares bigrams with FCUL_09229 protein. Therefore, it is highly possible that this FC protein might be involved in the pathogenicity process of FC.

7.4.2.4 Defining orphan proteins

In this section, the analysis is focused on the identification of proteins that are not part of PBON. As introduced in the earlier section of this chapter (see section 7.4.2.1), not all proteins with at least two domains participated in the construction of PBON. These include 16 FGRRES proteins, 44 FV proteins, and 260 FC proteins. As such, these proteins were assumed not to share sequence similarity with other *Fusarium* species in this study and could be considered to be orphan proteins, possibly species-specific. In Table 7-10 are listed FGRRES proteins that have at least two domains and are not part of the PBON. None of these proteins were found to be required for virulence, nor were predicted candidate genes by Lysenko et al. (2013), nor predicted as secreted proteins (Brown et al., 2012a).

Five FGRRES proteins, that have both pfam domains highlighted in bold in Table 7-10, were chosen for the further analysis. This is because these proteins only have pfam domains that are not present in FV and FC. Table 7-11 summarises the outcome of BLASTP searches using these five FGRRES proteins.

Table 7-10 *F. graminearum* proteins not incorporated into the PBON construction.

No	FGRRES protein Id	Total domains	Unique domains	Domains list			FGRRES transcript Id	Protein annotation*
1	CEF77191	3	1	PF00240	PF00240	PF00240	FGRRES_08768	probable UBI4 - ubiquitin
2	CEF77350	2	2	PF14529	**PF06839**		FGRRES_08889_M	related to APN2 - AP endonuclease, exonuclease III homolog
3	CEF84673	2	2	**PF01088**	PF00025		FGRRES_13502_3_M	probable ubiquitin thiolesterase L3
4	CEF74168	2	2	PF00240	**PF01020**		FGRRES_01956_M	probable ubiquitin fusion protein (ubiquitin / ribosomal protein)
5	CEF83185	2	2	**PF03732**	PF00098		FGRRES_20409	
6	CEF84502	2	2	PF00282	**PF01909**		FGRRES_13141	related to glutamic acid decarboxylase
7	CEF86734	2	2	**PF13446**	**PF00443**		FGRRES_12866	related to ubiquitin carboxyl-terminal hydrolase 2
8	CEF86541	2	2	PF00106	PF01370		FGRRES_04789_90_M	conserved hypothetical protein
9	CEF88719	2	2	PF00931	PF13414		FGRRES_11019	related to calcium-independent phospholipase A2
10	CEF87286	2	2	**PF04183**	**PF06276**		FGRRES_17587	hypothetical protein
11	CEF77420	2	2	PF05368	**PF01931**		FGRRES_13262_3_M	related to hypothetical protein yjjx
12	CEF85774	2	2	**PF13920**	PF14604		FGRRES_16806	hypothetical protein
13	CEF85014	2	2	**PF00149**	**PF13476**		FGRRES_06489_M	conserved hypothetical protein
14	CEF85394	2	2	**PF01336**	**PF09329**		FGRRES_09520	related to replication protein CDC23
15	CEF71927	2	2	**PF13401**	**PF03781**		FGRRES_11669	conserved hypothetical protein
16	CEF88685	2	2	**PF06172**	PF07690		FGRRES_12612_3_M	conserved hypothetical protein

*MIPS annotation. Highlighted in bold domains are domains that were found in FGRRES proteins not being part of the PBON and not found in either *F. culmorum* (FC) or *F. venenatum* (FV) proteins that are not part of the network.

Table 7-11 *F. gaminearum* proteins with two pfam domains highlighted in bold in Table 7-10.

No	FGRRES protein ID	FGRRES gene ID	Best BLASTP hit	Coverage [%]	Identity [%]
1	CEF86734	FGRRES_12866	*Fusarium pseudograminearum* CS3096 (HP FPSE_04241)	100	98
			Fusarium langsethiae (ubiquitin carboxyl-terminal hydrolase 2)	99	90
2	CEF87286	FGRRES_17587	*Fusarium graminearum* PH-1 (HP FGSG_11242)	100	91
			Fusarium graminearum (HP FG05_11242)	100	91
3	CEF85014	FGRRES_06489_M	*Fusarium graminearum*(HP FG05_30471)	100	82
			Fusarium graminearum PH-1 (HP FGSG_06489)	78	100
			Fusarium pseudograminearum CS3096 (HP FPSE_01771)	100	80
4	CEF85394	FGRRES_09520	*Fusarium graminearum* PH-1 (HP FGSG_09520)	100	99
			Fusarium pseudograminearum CS3096 (HP FPSE_01066)	100	99
			Fusarium langsethiae (minichromosome maintenance protein 10)	100	94
5	CEF71927	FGRRES_11669	*Fusarium oxysporum* FOSC 3-a (HP FOYG_00958)	100	78
			Fusarium oxysporum f. sp. *conglutinans race* 2 54008 (HP FOPG_13875)	100	77
			Fusarium oxysporum Fo5176 (HP FOXB_10588)	100	77

Where HP- hypothetical protein

In general, all these five proteins are only specific to the *Fusarium* genus. Furthermore, detailed investigation of Table 7-11 revealed that three proteins: CEF87286 (FGRRES_17587), CEF71927 (FGRRES_11669), and CEF85014 (FGRRES_06489_M) are specific to *F. graminearum*; whereas two proteins: CEF85394 (FGRRES_09520) and CEF86734 (FGRRES_12866) uncovered also a substantial similarity to proteins within *Fusarium pseudograminearum*.

7.4.3 Combined Protein Bigrams Overlap Network (CPBON)

The concept of PBON was intended to be utilised in the integration of protein sequences representing a wider taxonomical group of fungal species including both pathogenic and saprophytic lifestyles. This should help to find protein clusters specific to pathogenic and non-pathogenic fungi, as well as identify clusters specific to the different lifestyle of pathogenic fungi for example those attacking only leaves or roots of the plants. As stated above (subsection 7.3.3.2), three unrelated ascomycetes have been studied: *F. graminearum* (FGRRES) and *M. oryzae* (MO) and *N. crassa* (NC).

Here PBONs were constructed independently for each of three Ascomycota species. In FGRRES PBON 2,395 proteins (nodes) were connected via 26,351 edges and the network was spanned across 294 CCs. The number of nodes in MO PBON accounted for 1,477 with the connections between them equal to 13,498. The total number of CCs detected within MO PBON network was slightly lower when comparing to the number of CCs in FGRRES PBON and was equal to 270. In the PBON for NC 1,191 proteins were connected via 8,720 edges and the network consisted of 240 CCs.

In total 780 out of 3901 ortholog proteins (n1-n3901, Table 7-2) were in common for the above three PBONs, whereas 122 out of 938 ortholog proteins (n3902-n4839, Table 7-2) were common to the FGRRES and MO PBONs. Moreover, 98 out of 594 (n4840-n5433, Table 7-2) possible orthologs between FGRRES and NC were present in both their PBONs.

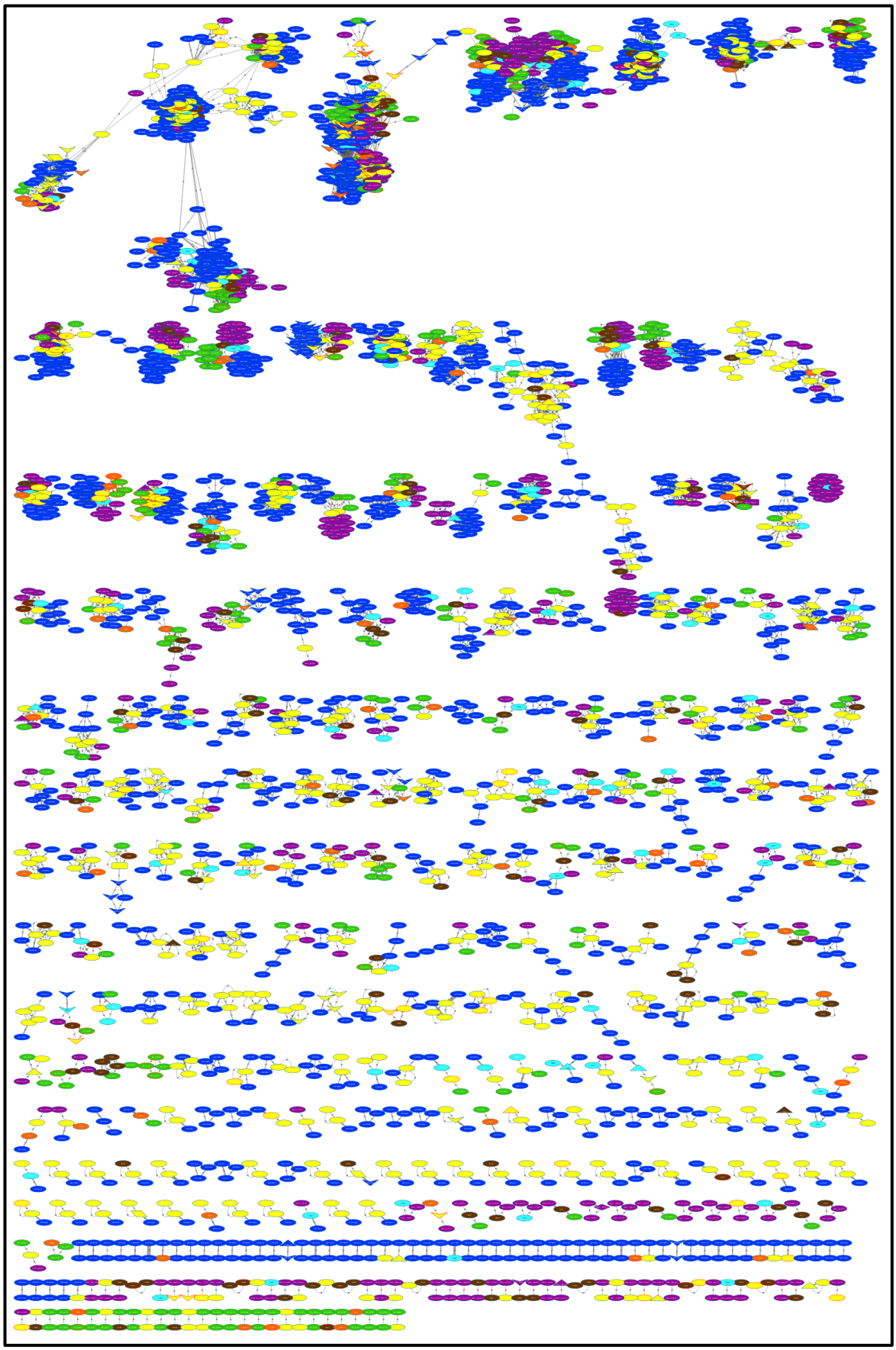
Finally, 173 out of 968 orthologs (n5434-n6401, Table 7-2) were in common for MO and NC. As a result of SimMod algorithm, three PBONs were clustered and connected via ortholog proteins

within at least two species. Consequently, a CPBON was created comprising in total 3,593 proteins connected via 48,572 edges. The 100 ‘composite’ modules (clusters) detected via SimMod were spanned across 376 components (Figure 7-6). Ninety six of these consisted only of proteins belonging to pathogenic. The majority (69) of the latter are represented by FG proteins only, whereas 14 were MO specific, leaving 13 CCs coming from both FG and MO, including orthologs of both species marked as cyan nodes (Figure 7-7).

Only five of the pathogen-specific components included a protein with associated phenotype in PHI-base version 3.8 (Urban et al., 2015b), see Figure 7-8. Detailed characteristic of the CCs from Figure 7-8 is summarised in Table 7-12.

The first Connected Component (CC) from Figure 7-8 demonstrates the connection between FGRRES and MO PBONs via ortholog proteins: CEF74812 (FGRRES_02506) and MGG_11343T0 (MGG_11343). The FG gene FG02506.1 of a SCK04 strain of FG (annotated as FGRRES_02506 of PH1 strain in the recent FG gene call (King et al., 2015)) has been shown to have an effect on the pathogenicity when testing on barley (Kim et al., 2007).

In that study, a mutant strain SCK04 of FG, generated by Restriction Enzyme-Mediated Integration (REMI), demonstrates multiple phenotypic changes such as reduction of mycelial growth, lack of sexual reproduction, as well as reduced virulence on barley. Further analysis of the REMI mutant strain in the study revealed that phenotypic changes were due to the mutation of the gene *ADE5* encoding for phosphoribosylamine-glycine ligase.



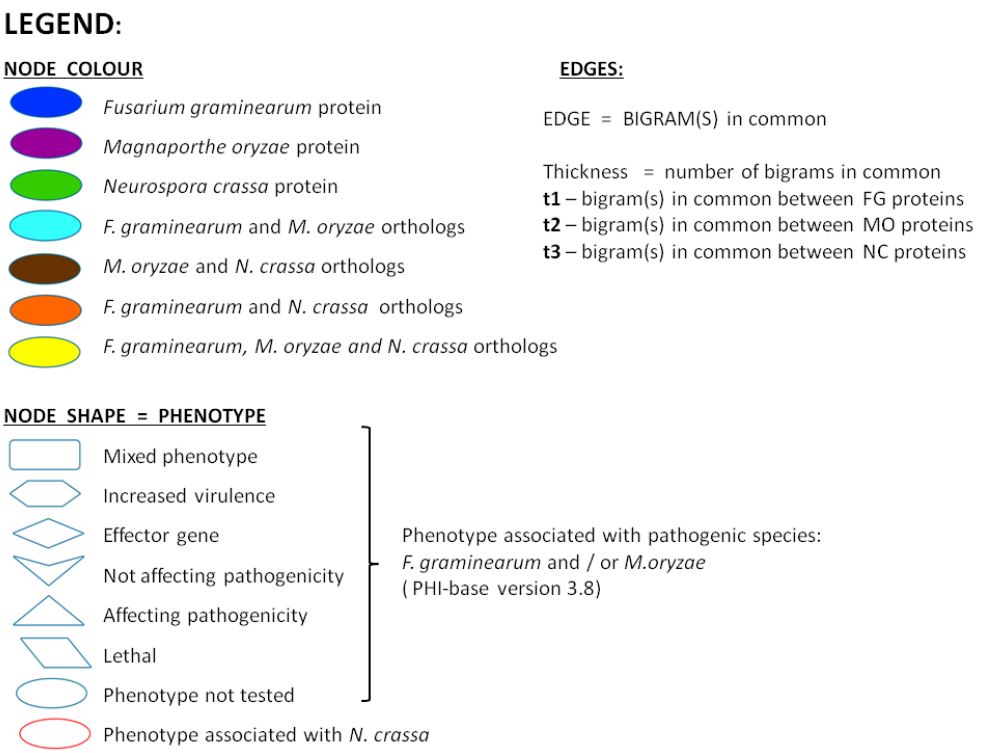
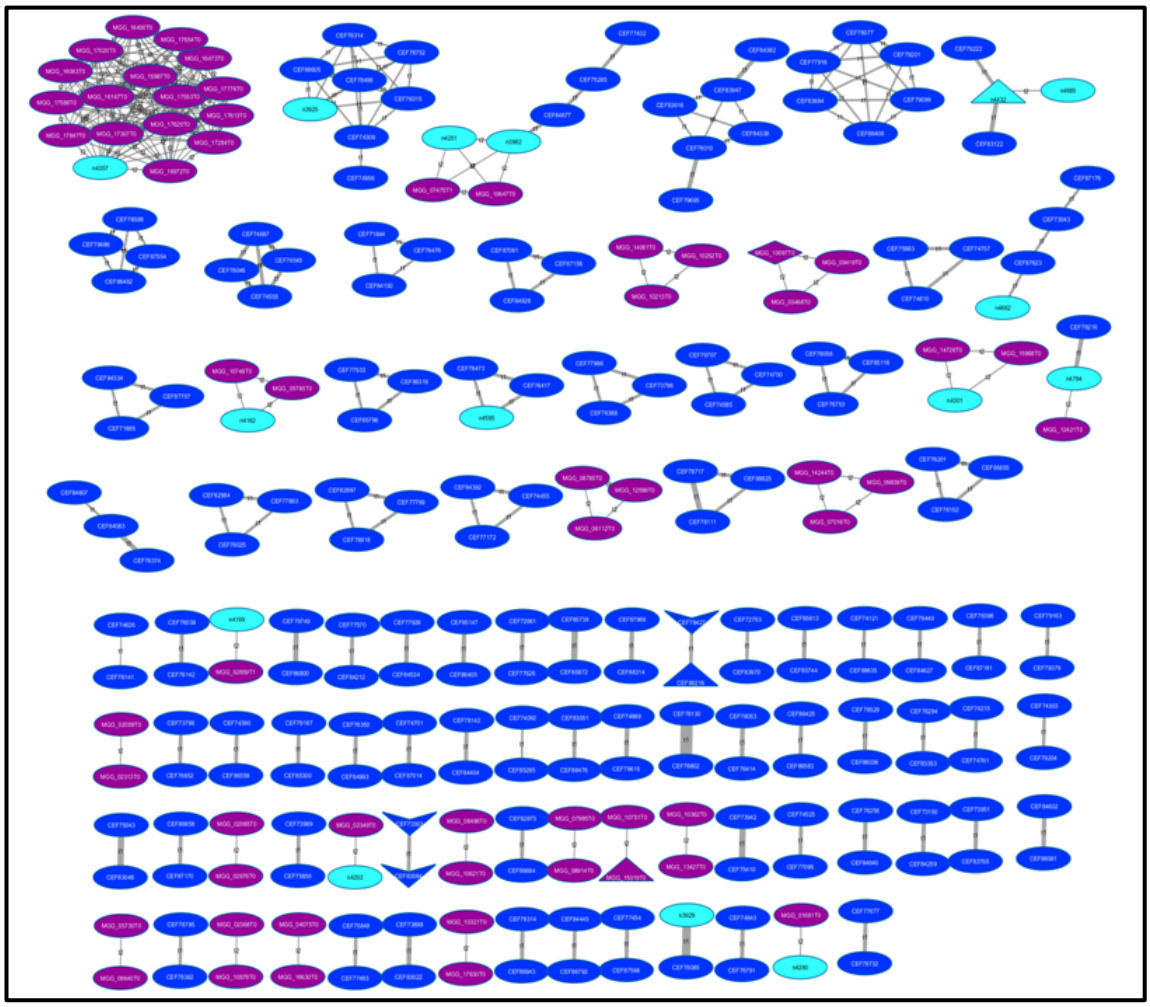


Figure 7-6 Combined Protein Bigrams Overlap Network (CPBON).
The network combines Protein Bigram Overlap Networks (PBONs) for *F. graminearum*, *M. oryzae* and *N. crassa* connected via orthologs between two or three species.



LEGEND:

NODE COLOUR

Fusarium graminearum protein

Magnaporthe oryzae protein

F. graminearum and *M. oryzae* orthologs

EDGES:

EDGE = BIGRAM(S) in common

Thickness = number of bigrams in common
t1 – bigram(s) in common between FG proteins
t2 – bigram(s) in common between MO proteins

NODE SHAPE = PHENOTYPE

Effector gene

Not affecting pathogenicity

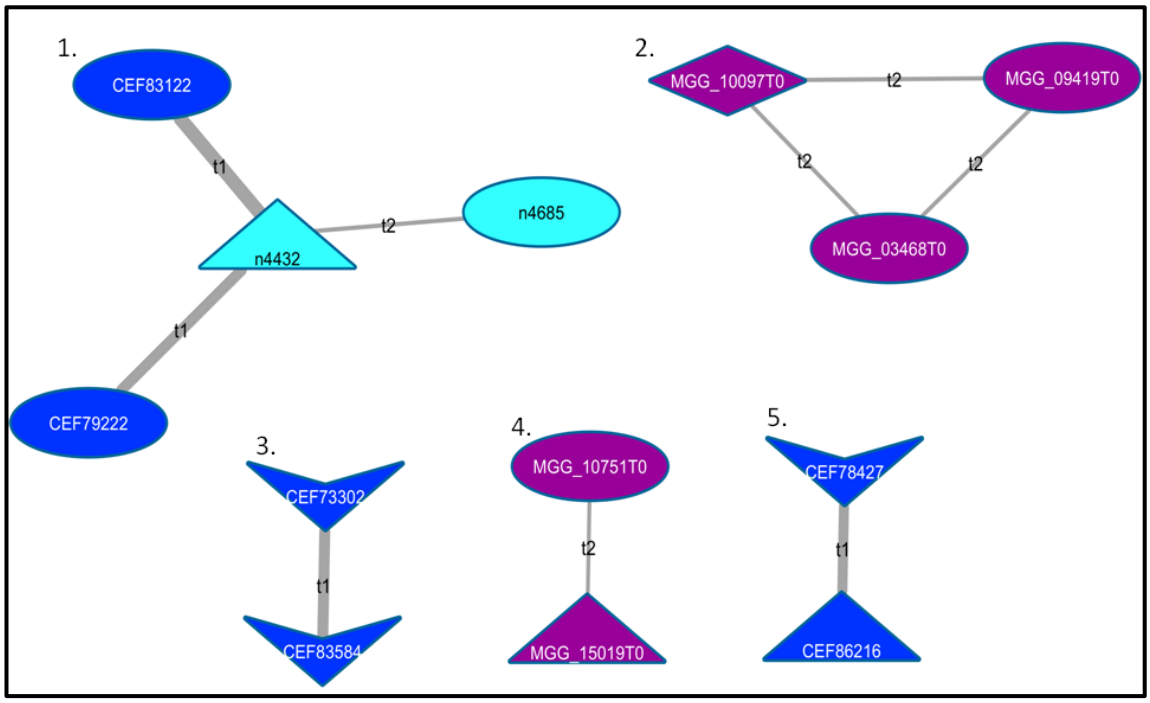
Affecting pathogenicity

Lethal

Phenotype not tested

Phenotype associated with pathogenic species:
F. graminearum and / or *M.oryzae*
(PHI-base version 3.8)

Figure 7-7 Pathogenic species connected components within Combined Protein Bigrams Overlap Network (CPBON).



LEGEND:

NODE COLOUR

Fusarium graminearum protein

Magnaporthe oryzae protein

F. graminearum and *M. oryzae* orthologs

EDGES:

EDGE = BIGRAM(S) in common

Thickness = number of bigrams in common
t1 – bigram(s) in common between FG proteins
t2 – bigram(s) in common between MO proteins

NODE SHAPE = PHENOTYPE

Effector gene

Not affecting pathogenicity

Affecting pathogenicity

Lethal

Phenotype not tested

Phenotype associated with pathogenic species:
F. graminearum and / or *M.oryzae*
(PHI-base version 3.8)

Figure 7-8 Subset of connected components with at least one protein having associated phenotype. Each component has been numbered. n4432 indicates orthologs between CEF74812 and MGG_11343T0, whereas n4685 indicates orthologs between CEF83122 and MGG_11541T0.

Table 7-12 Detailed characteristics of Connected Components illustrated in Figure 7-8.

CC No	Protein ID	FG gene ID	PHI-base phenotype*	PHI-base ID	Gene / function**	In common GO ***	GO term name / function
1	CEF83122	FGRRES_09440	Not tested			GO: 0003824 (5/5)	catalytic activity
	CEF79222	FGRRES_04250_M	Not tested			GO:0005737 (4/5)	cytoplasm
	CEF74812	FGRRES_02506	Reduced virulence	PHI:744	*ADE5* / **PGLA**	GO:0006189 (4/5)	'de novo' IMP biosynthetic process
	MGG_11343T0	MGG_11343	Not tested			GO: 0009113 (3/5)	purine nucleobase biosynthetic process
	MGG_11541T0	MGG_11541	Not tested			GO:0004637 (3/5)	phosphoribosylamine-glycine ligase activity
						GO:0005524 (3/5)	ATP binding
						GO:0046872 (3/5)	metal ion binding
2	MGG_10097T0	MGG_10097	Effector	PHI:2404	*SLP1*	GO:0016998 (3/3)	cell wall macromolecule catabolic process
	MGG_09419T0	MGG_09419	Not tested				
	MGG_03468T0	MGG_03468	Not tested		*SLP2*		
3	CEF73302	FGRRES_01214	Unaffected pathogenicity	PHI:1349	*GzC2H009* / **TF**	GO:0003676 (2/2)	nucleic acid binding
	CEF83584	FGRRES_09857	Unaffected pathogenicity	PHI:1414	*GzC2H081* / **TF**	GO:0046872 (2/2)	metal ion binding
4	MGG_15019T0	MGG_15019	Reduced virulence	PHI:2171	MGG_15019 / **CuAO**	GO:0046872 (2/2)	metal ion binding
	MGG_10751T0	MGG_10751	Not tested			GO:0016491 (2/2)	oxidoreductase activity
						GO:0055114 (2/2)	oxidation-reduction process
						GO:0005507 (2/2)	copper ion binding
						GO:0048038 (2/2)	quinone binding
						GO:0008131 (2/2)	primary amine oxidase activity
						GO:0009308 (2/2)	amine metabolic process
5	CEF78427	FGRRES_03597	Unaffected pathogenicity	PHI:1636	*FgFlbB* / **TF**	GO:0038032 (2/2)	termination of G-protein coupled receptor signaling pathway
	CEF86216	FGRRES_16620	Reduced virulence	PHI:1641 PHI:2434	*FgFlbA* / **TF**	GO:0035556 (2/2)	intracellular signal transduction

*PHI-base version 3.8, **author designated function of protein with assigned phenotype (PHI-base version 3.8), where TF –transcription factor, PGLA - phosphoribosylamine-glycine ligase activity, CuAO - Cu-amine-oxidase. *** numbers in brackets indicate the number of proteins with a given GO annotation over the total number of proteins in the given Connected Component (CC).

Table 7-12 reveals that three out of five proteins comprising the CC have been annotated phosphoribosylamine-glycine ligase GO term /function. These are CEF74812 (FGRRES_02506), CEF79222 (FGRRES_04250_M) and MGG_11343T0 (MGG_11343). Protein MGG_11343T0 is an ortholog of the protein CEF74812 encoded by the FG02506.1 gene in a PH1 strain of FG and both of them share the same pfam-domain repertoire and consequently all four bigrams in order. Hence, MGG_11343T0 and CEF79222 are likely to be considered as candidate.

The second CC, illustrated in Figure 7-8, comprises only MO proteins. One of these proteins, namely MGG_10097T0 (MGG_10097) has been recognised as an apoplastic effector protein named as a secreted LysM Protein1 (*SLP1*) (Mentlak et al., 2012). An apoplastic effector is an effector deployed in the space, known as an apoplast, within the fungal cell wall and the plasma membrane of the host (Mentlak et al., 2012). That study revealed that the MO *SLP1* gene binds chitin and was able to silence chitin-triggered immunity in the rice host and the same enabled the fungus to grow within the host.

Moreover, in this study it was confirmed that the protein encoded *SLP1* gene consists of two LysM (PF01476) domains. As expected, the other two MO proteins from the second CC (MGG_09419T0 and MGG_03468T0) also consist of three and two PF01476 LysM domains respectively. The study by Mentlak et al. named MGG_03468T0 as SLP2 and noted a strong sequence similarity to the *Cladosporium fulvum* effector gene Ecp6.

Thorough investigation of the protein sequences comprising the second CC (Figure 7-8), revealed that Cysteine residues comprise 4% of the total number of amino acids (aa) in MGG_10097T0 (protein size: 162 aa), 5% of aa in MGG_09419T0 (protein size: 437 aa) and 2% of aa in MGG_03468T0 (protein size: 285 aa).

BLASTP searches with all three MO proteins revealed that the MGG_10097T0 is exclusive to *M. oryzae* strain 70-15, MGG_09419T0 shares 99% identity (with 78% coverage) with a protein of *M. oryzae* strain P131, and MGG_03468T0 shares 98 and 95% identity (with 100% coverage in both cases) respectively with the intracellular hyphae protein 1 of *M. oryzae* strain P131 and intracellular hyphae protein 1 of *M. oryzae* strain Y34. This suggests that both MGG_09419T0

and MGG_03468T0 are likely to be putative effectors proteins whose interacting partners have not been detected yet.

The remaining CCs group proteins from the same species. The third and fifth comprise proteins from FGRRES, whereas the fourth are from MO. MGG_15019T0 in the latter group has been associated with pathogenicity (PHI-base version 3.8) when testing on rice and barley hosts (Tucker et al., 2010). The second protein in this CC has not been reported as pathogenic, but should tested because it shares the same GO annotations and pfam domains.

Furthermore, the third CC comprises FGRRES proteins only. Both have been assigned transcription factor function and they do not affect the pathogenicity of FG (Son et al., 2011).

Finally, the fifth CC illustrated in Figure 7-8 groups two proteins from FG, but these two proteins have a contradictory phenotypic outcome with respect to pathogenicity. Both proteins encode transcription factor genes and the phenotype of the mutants of these proteins have been determined in the study by Son et al.(2011). While the mutant of the protein CEF78427 (FGRRES_03597) did not show a change in the virulence, gene mutation of the protein CEF86216 (FGRRES_16620), resulted in significant reduction in virulence.

This finding was confirmed by a further study by Park et al. (2012), where CEF86216 (FGRRES_16620), encoded by the FgFlbA gene and annotated as a regulator of G protein signalling (RGS), was shown to effect on the pathogenicity when tested on wheat, and also to influence spore germination and mycotoxin production. The same study showed that mutation of FgFlbB, encoding CEF78427 (FGRRES_03597), led to a defect in conidia morphology but not to reduced virulence. CEF78427 and CEF86216 are connected by the bigram PF00610|PF00615, but the viulence of the latter might be due to its much larger size, 727 as opposed to 474 aa.

7.4.3.1 General characterisation of composite modules of combined Protein Bigrams Overlap Networks (PBONs)

As a result of the SimMod clustering, the number of clusters combining only proteins from pathogenic species increased from 96 CCs to 134 modules (clusters or communities), whereas the number of the modules include the previously mentioned 96 pathogen-specific CCs (Figure 7-9). The number of modules containing at least one protein with a PHI-base annotation increased from 5 (in Figure 7.8) to 13 (see Figure 7-10). Detailed characteristics of the most interesting modules from Figure 7-10 is summarised in the tables: Table 7-13 to Table 7-15. From Figure 7-10, modules 1 to 5 have already been described above and modules 6 and 11 are not worth considering further because they have many nodes of which only one has a PHI-base annotation.

Modules 7, 8 and 9 (Table 7-13), however, comprise transcription factors (Son et al., 2011). The majority of the proteins within these modules were experimentally tested for the phenotypic outcome and were associated with unaffected pathogenicity phenotype (PHI-base version 3.8, (Urban et al., 2015b)). The exceptions here are two proteins for which phenotype outcome has not been tested yet. These are CEF74868 (FGRRES_16014) protein from module 7 and CEF79545 (FGRRES_04606) protein from module 9 (Table 7-13). Moreover, one protein from the module 8 (Table 7-13), namely CEF84070 (FGRRES_16926_M) has been associated with a lethal phenotype (Son et al., 2011).

The last module in Table 7-13, module 10, catalogues two proteins encoding protein kinase genes (Wang et al., 2011a). These are CEF84118 (FGRRES_07121) and CEF72364 (FGRRES_00433) and both of them were associated with an unaffected pathogenicity phenotype (PHI-base version 3.8, (Urban et al., 2015b)). Interestingly two MO proteins are also present in module 10 (Table 7-13), but they have no PHI-base annotation. However, protein MGG_04790T0 is an ortholog of the FG protein CEF72364 which suggests that it is likely to encode a protein kinase gene. Additionally, all proteins within module 10 share six identical GO annotations.

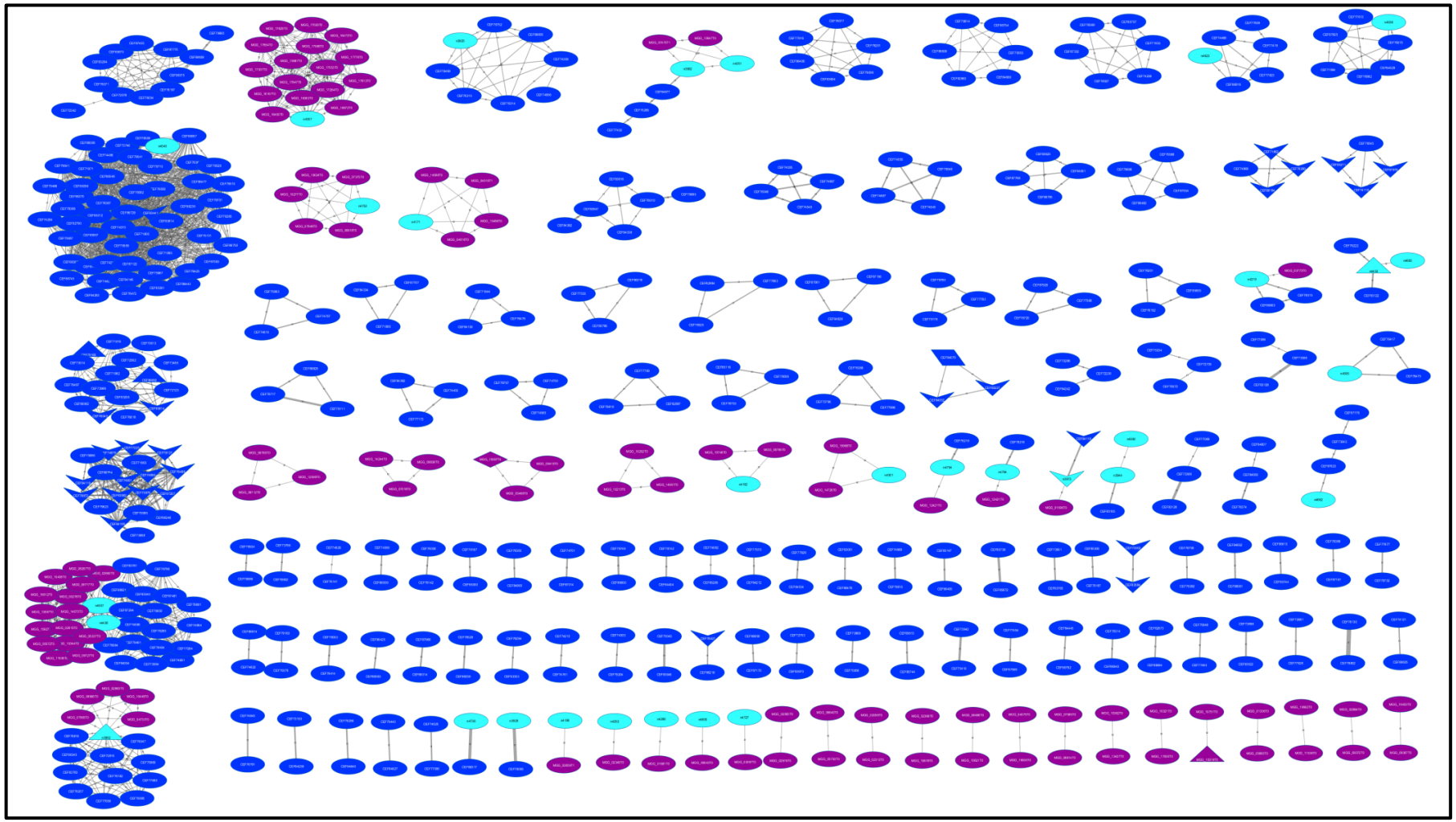
Further attention is focused on module 12 (Figure 7-10 and Table 7-14). It comprises of 16 FGRRES proteins, where two of them has been associated with a reduced virulence mutant phenotype and one of them, namely CEF79185 (FGRRES_04220) is a transcription factor (Son

et al., 2011), whereas the second one, namely CEF86652 (FGRRES_16412) is a protein kinase (Wang et al., 2011a). Two other proteins in this module, CEF76043 (FGRRES_17398) and CEF85614 (FGRRES_14027_16070_11614_M), are also protein kinases but according to PHI-base do not affect pathogenicity.

Finally, module 13 (Figure 7-10, Table 7-15) represents a cluster of 19 FG proteins where most of them (12 proteins) are multi-domains polyketide synthases (PKS) and were associated with mutant phenotypes not affecting pathogenicity.

Additionally, one of the proteins in this module, namely CEF73858 (FGRRES_15676), belongs to non-ribosomal peptide synthetases (NRPS). This protein has not been tested for the phenotypic outcome yet. Furthermore, four of PKS genes listed in Table 7-15 have been tested for functionality in *Fusarium* and their products comprise Aurofusarin (AUR1) (Gaffoor et al., 2005), Zearalenone (ZEA1, ZEA2) (Gaffoor et al., 2005, Kim et al., 2005, Lysoe et al., 2006) and Fusarin C (Gaffoor et al., 2005). The six unannotated proteins in this module are very likely to be PKS or NRPS enzymes.

The remaining modules include protein from *N. crassa* and were not studied further because their significance to pathogenicity is questionable, this would take a long time, and, hence, be unlikely to yield anything of interest.



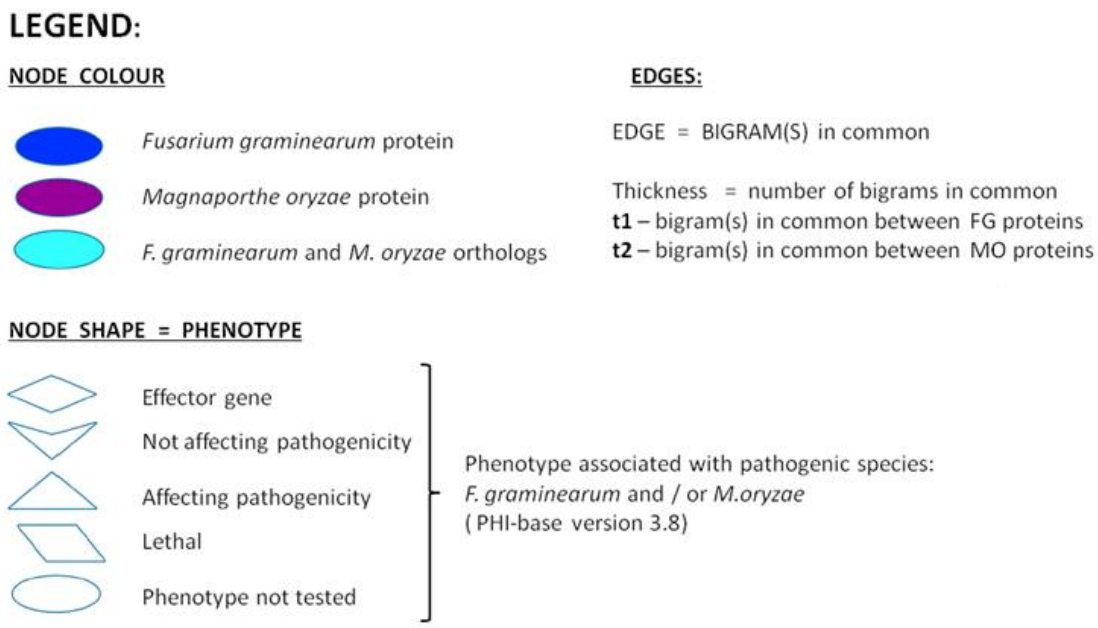
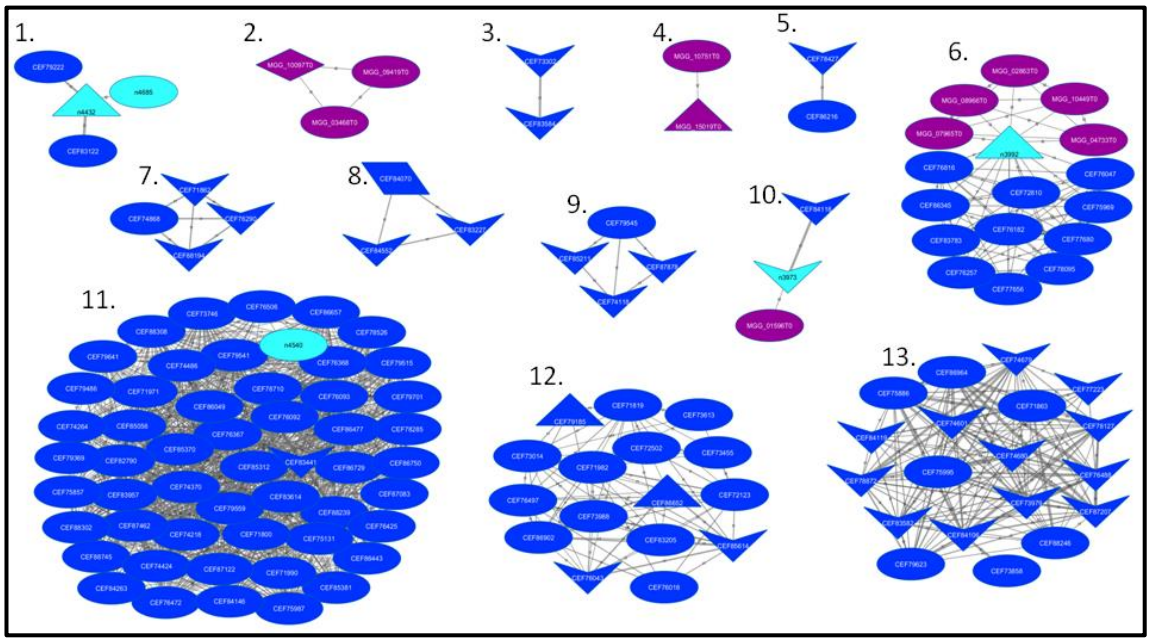


Figure 7-9 Pathogenic species modules (clusters) detected within Combined Protein Bigrams Overlap Network (CPBON).



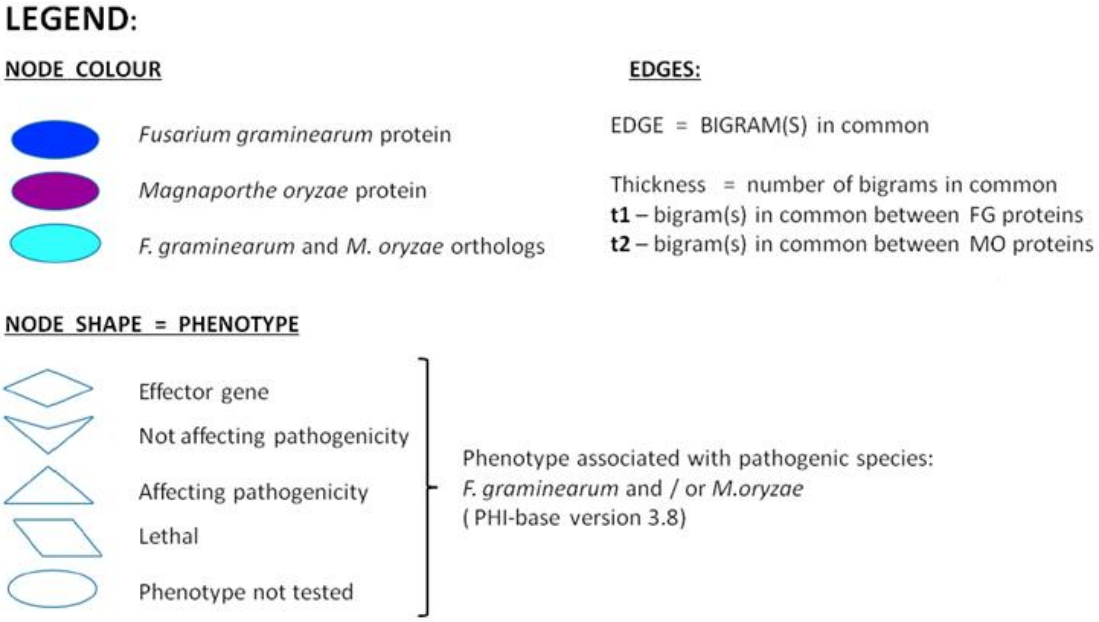


Figure 7-10 Pathogenic species modules in the Combined Protein Bigrams Overlap Network with associated phenotypes.
The modules have been numbered for clarity.

Table 7-13 Detailed characteristics of modules 7 to 10 illustrated in Figure 7-10.

Module No	Protein ID	Gene	PHI-base Phenotype*	PHI-base ID	Gene / function**	In common GO***	GO term name
7	CEF71862	FGRRES_11654	Unaffected pathogenicity	PHI:1750	GzZC065 / **TF**	GO:0008270 (4/4)	zinc ion binding
	CEF76290	FGRRES_08038	Unaffected pathogenicity	PHI:1803	GzZC118 / **TF**	GO:0006355 (4/4)	regulation of transcription, DNA-templated
	CEF74868	FGRRES_16014	Not tested			GO:0005634 (4/4)	nucleus
	CEF88194	FGRRES_05925	Unaffected pathogenicity	PHI:1878	GzZC193 / **TF**	GO:0003677 (4/4)	DNA binding
						GO:0006351 (4/4)	transcription, DNA-templated
						GO:0000981 (4/4)	RNA polymerase II transcription factor activity, sequence-specific DNA binding
						GO:0003700 (2/4)	transcription factor activity, sequence-specific DNA binding
8	CEF84070	FGRRES_16926_M	Lethal	PHI:1394	GzC2H057 / **TF**	GO:0006355 (3/3)	regulation of transcription, DNA-templated
	CEF84552	FGRRES_07914	Unaffected pathogenicity	PHI:1395	GzC2H058 / **TF**	GO:0005634 (3/3)	nucleus
	CEF83227	FGRRES_17150	Unaffected pathogenicity	PHI:1512	GzHOME010 / **TF**	GO:0003677 (3/3)	DNA binding
9	CEF79545	FGRRES_04606	Not tested			GO:0003677 (3/4)	DNA binding (FGRRES_04606, GO:0016787, hydrolase activity)
	CEF87878	FGRRES_13911	Unaffected pathogenicity	PHI:1528	GzHOMEL041 / **TF**	GO:0003682 (3/4)	chromatin binding
	CEF74118	FGRRES_01915	Unaffected pathogenicity	PHI:1541	GzFlbD / **TF**		
	CEF85211	FGRRES_09807	Unaffected pathogenicity	PHI:1551	GzMyb015 / **TF**		
10	CEF84118	FGRRES_07121	Unaffected pathogenicity	PHI:1257	FGSG_13509 / **PK**	GO:0005524 (4/4)	ATP binding
	CEF72364	FGRRES_00433	Unaffected pathogenicity	PHI:1256	FGSG_00433 / **PK**	GO:0005515 (4/4)	protein binding
	MGG_04790T0	MGG_04790	Not tested			GO:0016772 (4/4)	transferase activity, transferring phosphorus-containing groups
	MGG_01596T0	MGG_01596	Not tested			GO:0006468 (4/4)	protein phosphorylation
						GO:0004674 (4/4)	protein serine/threonine kinase activity
						GO:0004672 (4/4)	protein kinase activity

*PHI-base version 3.8, **author designated function of protein with assigned phenotype (PHI-base version 3.8), where TF –transcription factor, PK – Protein kinase. *** numbers in brackets indicate the number of proteins with a given GO annotation over the total number of proteins in the given module.

Table 7-14 Detailed characteristics of module 12 illustrated in Figure 7-10.

Module No	Protein ID	Gene	PHI-base Phenotype*	PHI-base ID	Gene / function**	In common GO***	GO term name
12	CEF79185	FGRRES_04220	Reduced virulence	PHI:1293	GzAPSES001 / **TF**	GO:0005515 (16/16)	Protein binding
	CEF86652	FGRRES_16412	Reduced virulence	PHI:1209	FGSG_04770 / **PK**	GO:0005524 (3/16)	ATP binding
	CEF76043	FGRRES_17398	Unaffected pathogenicity	PHI:1283	FGSG_10591 / **PK**	GO:0006468 (3/16)	Protein phosphorylation
	CEF85614	FGRRES_14027_16070_11614_M	Unaffected pathogenicity	PHI:1280	FGSG_11614 / **PK**	GO:0004674 (3/16)	Protein serine/threonine kinase activity
	CEF71819	FGRRES_11644	Not tested			GO:0004672 (3/16)	Protein kinase activity
	CEF71982	FGRRES_15694	Not tested				
	CEF72123	FGRRES_15713	Not tested				
	CEF72502	FGRRES_00550	Not tested				
	CEF73014	FGRRES_00965	Not tested				
	CEF73455	FGRRES_01338_M	Not tested				
	CEF73613	FGRRES_15852	Not tested				
	CEF73988	FGRRES_01799_M	Not tested				
	CEF76018	FGRRES_10569	Not tested				
	CEF76497	FGRRES_17083	Not tested				
	CEF83205	FGRRES_09129	Not tested				
	CEF86902	FGRRES_11372	Not tested				

*PHI-base version 3.8, **author designated function of protein with assigned phenotype (PHI-base version 3.8), where TF –transcription factor, PK – Protein kinase. *** numbers in bracket indicate the number of proteins with a given GO annotation over the total number of proteins in the given module.

Table 7-15 Detailed characteristics of module 13 illustrated in Figure 7-10.

Module No	Protein ID	Gene	PHI-base Phenotype*	PHI-base ID	Gene / function**	In common GO***	GO term name
13	CEF73979	FGRRES_01790	Unaffected pathogenicity	PHI:725	PKS11 / **PS**	GO:0008152 (19/19)	Metabolic process
	CEF74601	FGRRES_02324	Unaffected pathogenicity	PHI:720	PKS12 AUR1/ **AS**	GO:0003824 (19/19)	Catalytic activity
	CEF74679	FGRRES_15980_M	Unaffected pathogenicity	PHI:713	PKS13 ZEA2 / **ZS**	GO:0016740 (17/19)	Transferase activity
	CEF74680	FGRRES_17745	Unaffected pathogenicity	PHI:714	PKS4 ZEA1 / **ZS**	GO:0031177 (13/19)	Phosphopantetheine binding
	CEF76488	FGRRES_08208	Unaffected pathogenicity	PHI:718	PKS6 / **PS**		
	CEF77223	FGRRES_08795	Unaffected pathogenicity	PHI:726	PKS7/ **PS**		
	CEF78127	FGRRES_03340	Unaffected pathogenicity	PHI:723	PKS1 / **PS**		
	CEF78872	FGRRES_03964	Unaffected pathogenicity	PHI:721	GRS1/ **PS**		
	CEF83582	FGRRES_17168	Unaffected pathogenicity	PHI:722	PKS3 PGL1 / **BpEPS**		
	CEF84106	FGRRES_07798	Unaffected pathogenicity	PHI:719	PKS10 GzFUS1/ **FCS**		
	CEF84119	FGRRES_07226_M	Unaffected pathogenicity	PHI:724	PKS9/ **PS**		
	CEF87207	FGRRES_04694	Unaffected pathogenicity	PHI:729	PKS2 / **PS**		
	CEF71863	FGRRES_00036	Not tested				
	CEF73858	FGRRES_15676	Not tested		NPS19		
	CEF75886	FGRRES_10464	Not tested				
	CEF75995	FGRRES_17387	Not tested				
	CEF79623	FGRRES_04588	Not tested				
	CEF86964	FGRRES_17677	Not tested				
	CEF88246	FGRRES_05321	Not tested				

*PHI-base version 3.8, **author designated function of protein with assigned phenotype (PHI-base version 3.8), where PS – Polyketide synthase, AS – Aurofusarin synthesis, ZS – Zearalenone synthesis, BpEPS – Black perithecial pigment synthesis, FCS – Fusarin C synthesis. *** numbers in bracket indicate the number of proteins with a given GO annotation over the total number of proteins in the given module.

7.5 Discussion

In this chapter, various types of network analyses were attempted to predict or elucidate unknown function of proteins in plant pathogenic fungi. In general, three different approaches were employed towards achievement of this goal.

The first approach involved prediction of host pathogen protein-protein interactions (HPPPI) for the two economically important plant pathogenic fungi, namely *F. graminearum* and *M. oryzae*, using rice as their host. Unfortunately, the initial investigation revealed that it would be not possible to use the PPIN-1 system (Mukhtar et al., 2011) to validate the prediction.

Consequently, following the unsuccessful approach to predict and validate the HPPPI it was then interesting instead to examine interactions between proteomes of fungi representing different lifestyles including both pathogenic and saprophytic fungi. Firstly, a Protein Bigrams Overlap Network (PBON) concept was employed to integrate proteins from three closely related species of the same genus. Two of these species, namely *F. graminearum* and *F. culmorum* represent plant pathogenic fungi, whereas *F. venenatum* is a non-pathogenic saprophyte fungus.

As expected, overall participation of proteins from all three species in the PBON network was almost uniform. This was very well visualised in Figure 7-5, where most CCs are comprised of proteins of all the above three species. In addition, statistical tests confirmed that the biological lifestyle of the three *Fusarium* species did not affect the distribution of the node (protein) parameters within the network. Therefore, the subsequent analysis focused on the smallest CCs in order to identify pathogenic-lifestyle and species-specific clusters in the PBON.

In total 28 CCs catalogued proteins from only pathogenic species and three of these CCs were specific to FC. Thus, 25 CCs with two nodes grouped proteins from both FGRRES and FC. Moreover, most of the FGRRES proteins in these CCs are very well annotated (see Table 7-7) with the exception of nine hypothetical proteins. Although in the majority of these clusters proteins share only one pfam domain bigram, the FGRRES protein annotation may be helpful in finding the annotation for paired FC proteins.

In addition, 16 orphan FGRRES proteins were distinguished within this study. Five of these proteins contain pfam domains bigrams that were only found in FGRRES and not in the other two species.

BLASTP searches revealed that three of them are *F. graminearum* specific: CEF87286 (FGRRES_17587), CEF85014 (FGRRES_06489_M) and CEF71927 (FGRRES_11669). Not surprisingly all of these proteins are hypothetical proteins. Furthermore, protein CEF71927 (FGRRES_11669) has been found in the highest recombination region on the chromosome I of FGRRES genome and no highly similar protein sequence (with 100% coverage and identity at or above 95%) was found using BLASTP.

The concept of PBON application and use of SimMod have been extended to integrate proteomes from pathogenic and saprophytic ascomycetes, specifically *F. graminearum*, *M. oryzae* and *N. crassa.*

The main attention in the final analysis was focused on the clusters representing proteins from the pathogenic species. Several biologically interesting modules were discovered. These are depicted in Figure 7-10 and include modules with *M. oryzae* effectors genes (module 2), PKS/ NRPS enzymes (module 13), transcription factors (modules 7, 8 and 9) and proteins kinases (modules 10 and 12). In addition, several assumptions about possible phenotypic outcome of the proteins in the detected modules were made. This of course needs further investigation. All genes implicated in this study are putative targets that can be tested in gene deletion experiments for example. Furthermore, RNA expression of these genes under disease-causing condition (during the host infection) can be investigated.

In general, the final concept in this chapter demonstrates the successful adaptation of the previously developed method (Bennett et al., 2015) into a new analysis where biological information from several different species, representing different lifestyles, has been integrated into one platform. Then, using several known functional protein annotations and known phenotypic outcomes, it has become possible to elucidate the function for unknown gene or proteins in the network.

Chapter 8

General discussion

The main motivation of this thesis was to explore proteins of unknown function, as well as speculate or even predict their role in the sequenced genomes of plant pathogenic fungi. Several different network approaches were employed in order to use and integrate the available biological data to assign new annotations.

Among the new approaches delivered in this work was the application of the domain-association network in function prediction of DUF. Surprisingly, DUFs appeared not to be directly linked to the pathogenicity of fungal plant pathogens. Moreover, DUFs were found to comprise the peripheral nodes in domain-association network modules and as such did not interact with nodes of other communities. Additionally, further study of the domain network topologies revealed DUF association with lethal proteins suggesting that they could be linked to *F. graminearum* essential proteins. The PPI network analysis in Chapter 4 led to candidate gene prediction for the plant pathogenic fungus *F. graminearum* and 12 of them were later experimentally validated as the ones required for the pathogenicity of *F. graminearum* (PHI-base version 3.8) (Urban et al., 2015b).

Domain bigram, domains' network, and bigram network approaches were earlier introduced into the bioinformatics analysis by previous studies (Seidl et al., 2011, Wang et al., 2011c, Xie et al., 2011). However, the work presented in this thesis is novel in the way that it is focused mainly on the prediction of DUF roles in plant pathogenic fungi. By integrating biological data such as protein predicted phenotype, taxonomic diversity of domain, as well as topological properties of domains within the network, it was possible to anticipate the role of DUFs within fungal plant pathogens. It was suggested that DUFs are frequently associated with essential / lethal proteins of *F. graminearum*. That is why their function is unknown. Also, in this study, a method for solving the overlapping of domains within the protein was modified and successfully implemented in Chapters 5, 6, and 7.

8.1 Overview of the thesis

Firstly, in Chapter 3, the network methodology was used to generate sequences similarity clusters of proteins / genes included in PHI-database versions 3.2 and 4.0. As a result, plant pathogens, animal pathogens as well as other host pathogens clusters were identified.

Then, in Chapter 4, a PPI network was employed to predict candidate genes for pathogenicity in a single fungal species. Subsequently, in Chapter 5, a bigram approach was used to construct a domain-association network for exploring of DUFs characteristics.

Finally, in Chapter 7, a series of PBONs were built for multiple species, starting with closely related *Fusarium* species, comprising both pathogenic and non-pathogenic representatives. Then, a previously developed method (Bennett et al., 2015) was successfully adapted into the analysis of the same type of interactions, namely PBONs constructed from three species across Ascomycota fungi with diverse biological lifestyles. This analysis further revealed several genes which need additional laboratory investigation for their effect on pathogenicity:

- MGG_09419 and MGG_03468 – novel effector protein candidates in MO. They are crucial during fungal-plant interactions to control, disable the host immune system and facilitate colonization of fungi.
- FGSG_00036, FGSG_04588, FGSG_05321, FGSG_10464, FGSG_17387 and FGSG_17677 (FG) – members of multi-modular enzymes: PKS and NRPS. Their known products include bioactive secondary metabolites (mycotoxins) that lead to health problem if consumed by animals or humans.

8.2 Research aims revisited

In Chapter 1, four research aims were outlined. These are now revisited in this section to determine how successfully they have been fulfilled.

- *To identify plant pathogen-specific gene clusters and animal pathogen-specific gene clusters required for virulence, as well as those required by both pathogen types.*

This aim was addressed in Chapter 3. The network methodology was used to generate sequences similarity clusters of proteins / genes included in PHI-database versions 3.2 and 4.0. In addition, the content of both versions of PHI-base was compared. As a result, clusters that contain genes associated with plant, animal, and both types of pathogens were identified in both versions of PHI-base. Therefore, this research aim was achieved.

- *To identify and predict the pathogenicity gene complement of two economically important plant pathogenic fungi, namely Fusarium graminearum and Magnaporthe oryzae*

This goal was tackled in Chapter 4. Previously predicted PPI networks were employed to predict candidate genes for both economically important plant pathogens (He et al., 2008, Zhao et al., 2009). The analysis in this chapter led to the prediction of 65 *F. graminearum* candidate genes for the plant pathogenic fungus *F. graminearum* and 12 of them were later experimentally validated as the ones required for the pathogenicity of *F. graminearum* (PHI-base version 3.8) (Urban et al., 2015b).

Unfortunately, using the network approach described in this chapter, it was not possible to predict all candidate genes for pathogenicity in *F. graminearum.* This is because some important species-specific genes, contributing to pathogenic lifestyle of *F. graminearum,* could not be mapped to the FPPI networks used in this study. The analysis for *Magnaporthe oryzae* was not expanded because of the low number of predicted candidate genes for pathogenicity. Consequently, the first part of this research aim was achieved.

- *To investigate the role of Domains of Unknown Function (DUF) in the pathogenicity of the plant pathogenic fungus Fusarium graminearum*

This research aim was addressed in Chapters 5 and 6. The first step was tackled in Chapter 5, where a combined strategy was applied, incorporating *F. graminearum* pfam domain repertoire identification, as well as *F. graminearum* pfam domain diversity evaluation, and finally use of a bigram approach (Seidl et al., 2011) to construct a domain-association network for exploring of DUFs characteristics. Three DUFs were initially associated with pathogenic lifestyle of *F. graminearum*: DUF619, DUF3546 and DUF4187. The first of them was the only domain in

FGSG_01939 protein and the protein was experimentally proven to be required for virulence (PHI-base versions 3.4 and 3.6 (Urban et al., 2015b)). The remaining two DUFs contribute to the only bigram (two domains) in FGSG_01106 protein, which was also experimentally proven to be required for virulence (PHI-base versions 3.4 and 3.6 (Urban et al., 2015b)).

Additionally, the taxonomical study of DUFs identified 35 fungal-specific DUFs. These DUFs, however, did not include three DUFs identified in the experimentally proven virulent FG proteins, namely FGSG_01939 and FGSG_01106. Furthermore, chi-square tests revealed that only DUF3129 has a stronger association with plant pathogenic fungi.

In the analysis conducted in Chapter 6, DUF content and diversity were compared between plant fungal pathogens such as *F. graminearum* and *F. culmorum*, and non-pathogenic fungi also representing *Fusarum* genus, namely *F. venenatum*. Amongst these species the same, as per Chapter 5, 35 fungal-specific DUFs were identified. Moreover, these DUFs were almost evenly distributed throughout proteomes of studied *Fusaria* in Chapter 6.

Based on the analyses conducted in Chapters 5 and 6 it is difficult to speculate the direct role of DUFs in the pathogenicity of *F. graminearum.* It is therefore concluded that this research goal was met confirming a negative outcome of the investigation as DUFs appeared not to be directly linked to the pathogenicity of fungal plant pathogens.

- *To perform a comparative network-based study between closely-related, as well as more distantly-related Ascomycetes, including both pathogenic and non-pathogenic fungi to reveal novel insights into pathogenicity.*

This research goal was addressed in Chapter 7. Various types of network analyses were attempted to predict or elucidate unknown function of proteins in plant pathogenic fungi. Initially, a HPPPI network concept was investigated to identify protein functions vital for pathogenicity in both economically important plant pathogenic fungi, namely *F. graminearum* and *M. oryzae*. However, due to the lack of validation method of possible HPPPI prediction the initial concept was abandoned. Consequently, a series of PBONs were built, starting with closely related *Fusarium* species, comprising both pathogenic and non-pathogenic representatives. Then, a

previously developed method (Bennett et al., 2015) was successfully adapted into a new approach to integrate the same type of interactions, namely PBONs constructed from three species across Ascomycota fungi with diverse biological lifestyles.

As a result of the investigation conducted in Chapter 7, *F. graminearum*-specific proteins were identified. Moreover, several biologically interesting protein modules were discovered within the pathogenic species. These include effector genes in *M. oryzae*, as well as PKS/ NRPS enzyme and transcription factor modules identified in *F. graminearum*. In addition, several assumptions about possible phenotypic outcome of the proteins in the detected modules were made. The outcome of this study demonstrates that the application of different concepts of PPI networks helps to some extend to reveal novel insight into fungal pathogenic lifestyle. It is therefore concluded that this research goal was met although further investigation would need to be conducted to confirm the findings of this study.

8.3 Contributions of the thesis

Key contributions of this thesis are listed below:

- There exist common, as well as unique pathogenicity determinants for plant and animal species;
- Cereal invading filamentous fungal species have a unique gene repertoire which enabled them to become successful plant pathogens;
- Prediction of candidate pathogenic genes in *F. graminearum* and the co-authored publication (Lysenko et al., 2013);
- There exist DUFs specific to fungi;
- Proteins with only domain that is DUF are essential proteins in fungal species;
- Implementation and adaptation of the SimMod method (Bennett et al., 2015) into a novel analysis.
- 9 genes were suggested for further laboratory experiments. These include genes listed in section 8.1, as well as *F. graminearum* gene (FGSG_06444) identified in Chapter 4 as a potential candidate gene responsible for mycotoxin production.

8.4 Handled difficulties

During the course of this study, several difficulties were encountered. PHI-base content has vastly expanded. In addition, *F. graminearum* genome release versions and annotation changed frequently leading to newly called genes in 2015 (King et al., 2015). In Chapter 4, initially PHI-base version 3.2 was used alongside *F. graminearum* gene assembly 3, whereas in the latest Chapter 7, the PHI-base version 3.8 and the fully completed *F. graminearum* genomic sequences with a refined proteome (King et al., 2015) was used. Moreover, while concluding this work in December 2015 the PHI-base version 4.0 became available. Thus, this version of PHI-base has been implemented into Chapter 3 and its size and content compared to the PHI-base version 3.2. The differences between the two *F. graminearum* proteomes were investigated in Chapter 6 of the thesis.

Furthermore, PFAM domain database releases occurred annually, which led to different content of DUFs within different versions of PFAM. This is due to the fact that every year new DUFs are added into the database, and some DUFs from the previous database version are either renamed with an identified function, merged with other pfam domains, or completely deleted from the database. However, the changes in the PFAM did not affect the study in this thesis, as throughout the whole of this study PFAM version 27.0 was used.

While concentrating on this project, several additional resources became available. This includes Ensembl Fungi (Kersey et al., 2014), PhytoPath (Pedro et al., 2016), as well as *Neurospora crassa* phenotypic data (from BROAD).

8.5 Future work

Future work will include the improvement and extension of the approaches presented in this thesis. As there has not been further development in the prediction of fungal plant pathogen interactomes since 2009, future work would mainly concentrate on the prediction of new protein-protein interaction (PPI) network for the newly released *F. graminearum* genome assembly and associated modified proteome (FG PPI) (King et al., 2015).

The interactome would be constructed using two established methods, namely the interolog approach and domain-domain interaction approach. The interolog approach requires reference interactome(s) and mapped ortholog sequences that could link it to a species of interest. *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* can be used as two reference interactome species. This is because both have some of the best-profiled, experimentally verified interactomes and both are phylogenetically very close to the pathogenic target fungi. Interacting data for these two species will be taken from the EBI Intact database (Orchard et al., 2014). Data will then be combined with orthologs retrieved from Ensembl Fungi (Kersey et al., 2018), which were originally derived using the EnsemblCompara pipeline (Herrero et al., 2016).

A prerequisite for the domain-domain interaction (DDI) approach is that some of the interactions are mediated by specific protein domains. It can therefore be assumed that these interactions occur between the paired proteins. Here, to obtain a more complete dataset, three domain-domain interaction databases: KBDOCK (Ghoorah et al., 2014), 3did (Mosca et al., 2014) and DOMINE (Yellaboina et al., 2011) could be used. The domain repertoire for FG will be identified using HMMER algorithm and domain models available in the latest PFAM database. Additionally, the overlapping of domains in the same protein will be resolved as per methodology implemented initially in Chapter 5. Then, this non-redundant domain dataset in FG will be used to infer interactions for each pair of proteins containing interacting domains included in at least one of the three DDI databases.

To verify the quality of different sources of inferred interactions, summary statistics will be calculated for the number of predicted interacting partners found in the same cellular compartment and having functional similarity per Gene Ontology (GO) annotation in biological process (BioP) and molecular function (MolF) aspects. The expected pattern is that true positive interactions would be found in the same compartment and be functionally similar. This of course would need to be further validated by comparing to *S. cerevisae* experimental interactions for example.

Furthermore, available transcriptomic data for FG could be used to construct an expression RNAexp (FG RNAexp) network as previously described in Lysenko et al., (2013). In addition, another PPI network could be constructed based on the ordered bigram similarity of the proteins

in the network (PBON), as implemented in Chapter 7. Then, three networks: FG PPI network, FG RNAexp network and PBON could be connected via FG ids and composite modules would be detected using SimMod network clustering. Consequently, further analyses on the composite modules will be performed. These would include: mapping of FG predicted phenotypes from the latest PHI-base database into the modules of combined FG network, using enrichment analysis method to predict candidate genes for pathogenesis and prioritizing genes based on module membership and overall score of the module generated by SimMod.

Future work will also involve improvement to the prediction of pathogenic genes in fungi via the application of different statistical and computational methods. The improvement to the prediction of disease causing genes in plant pathogenic fungi could be achieved by application of machine learning approaches in biological network analysis. Despite considerable advances, previous studies have been subject to considerable limitations, chiefly due to insufficient manually curated data being available to precisely quantify the performance of different proposed methods (Lysenko et al., 2013). Therefore, the validation was typically based on expert review, indirect evidence, small-scale biological validation or concentrating on narrow groups of genes like effectors, which have a highly specific set of properties (Sperschneider et al., 2015).

It would be particularly interesting to explore ways of lifting the limitations of the methods based solely on guilt-by-association (Gillis and Pavlidis, 2012) and to identify a set of predictive characteristics that can work in wide variety of fungal species and could still be applied when little or no phenotypic annotation and limited experimental data are available. Two different approaches could be potentially explored to relax or lift such limitations, namely random walk with restart (RWR) algorithm (Köhler et al., 2008) and machine learning based on a set of features indicative of gene importance in the network. The RWR algorithm works by quantifying a probability of a given node to be visited by a ‘walker’ starting from one of the known ‘seed’ nodes and randomly traversing edges in a given network. The RWR algorithm will be run using either sets of known pathogenicity related or pathogenicity unrelated genes. The results will be combined using random forest algorithm (Breiman, 2001) to produce an overall score. Here both pathogenicity related and pathogenicity unrelated subsets of nodes may be relevant for predicting

the phenotype, as random ‘walker’ starting from each of those subsets is on average more likely to visit the members of the same respective subset.

Adding a number of additional function annotations to PHI-base genes within the networks could also be considered. One set of features could comprise different measures of gene importance in inferred protein-protein interaction networks by including additional mutant phenotype data such as effect on growth in vitro, sporulation and sexual reproduction. These phenotypes often result in reduced pathogen survival under field conditions and can be independent from the virulence phenotype observed on hosts. The selected features will be used in combination with machine learning to identify the most promising candidate targets.

To support the model development, annotated genes would be divided into two subsets. The first of them would include fungal species for which the most annotation in PHI-base is available and will be used for feature selection and parameter optimization of the machine learning algorithm. In the second set all other fungal species, with at least one annotation in PHI-base, would be combined and used as an independent test set to validate the prediction method (model). The main limitation to this approach is that it relies on analysis of PPI networks to estimate the likely importance of genes. Both coverage and quality of a PPI network can be a limiting factor.

However, other potentially informative sources of knowledge that can be used in network construction are transcriptomic data and the metabolic pathway networks. Although the transcriptomics approach can be very informative, data are often not available in sufficient quantities for some of the key fungal pathogens. Inclusion of metabolic pathways map are unlikely to improve network coverage due to the fact that metabolomics data for pathogen-host interactions are scarce. However, additional links between host and pathogen can be of great importance to help with identification of candidate genes.

8.6 Concluding remarks

Network analysis plays an important role in the answering biological questions, as well as in defining the connections between biological systems. Several network approaches were

implemented throughout this study to combine available information, as well as further explore, investigate, or predict the role and function of proteins of interest.

This thesis has advanced the use and adaptation of different concept to elucidate, speculate, or even predict the unknown function of proteins in plant pathogenic fungi. These include bigram analysis (Seidl et al., 2011), Protein Overlap Network construction (Liang et al., 2013), as well as the SimMod algorithm (Bennett et al., 2015). Special attention is given to the SimMod method that was successfully implemented and adopted into a novel analysis where biological information from species representing different lifestyles can be integrated into the one platform. Then, using several known functional protein annotations and /or phenotypic outcomes, I was able to speculate on the function of unknown genes or proteins in the network.

Bibliography

AGRAWAL, Y., KHATRI, I., SUBRAMANIAN, S. & SHENOY, B. D. 2015. Genome Sequence, Comparative Analysis, and Evolutionary Insights into Chitinases of Entomopathogenic Fungus *Hirsutella thompsonii*. *Genome Biology and Evolution,* 7**,** 916-930.

AGRIOS, G. N. 2005. *Plant Pathology*, Academic Press.

AGUILAR-PONTES, M. V., DE VRIES, R. P. & ZHOU, M. 2014. (Post-)Genomics approaches in fungal research. *Briefings in Functional Genomics,* 13**,** 424-439.

ALBERT, R. & BARABÁSI, A.-L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics,* 74**,** 47-97.

ALBERT, R., JEONG, H. & BARABASI, A.-L. 2000. Error and attack tolerance of complex networks. *Nature,* 406**,** 378-382.

ALTAF-UL-AMIN, M., AFENDI, F. M., KIBOI, S. K. & KANAYA, S. 2014. Systems Biology in the Context of Big Data and Networks. *BioMed Research International,* 2014**,** 11.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25**,** 3389-402.

ANDREEVA, A., HOWORTH, D., CHANDONIA, J.-M., BRENNER, S. E., HUBBARD, T. J. P., CHOTHIA, C. & MURZIN, A. G. 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research,* 36**,** D419-D425.

ANTONIW, J., BEACHAM, A., BALDWIN, T., URBAN, M., RUDD, J. & HAMMOND-KOSACK, K. 2011. OmniMapFree: A unified tool to visualise and explore sequenced genomes. *BMC Bioinformatics,* 12**,** 447.

APIC, G., GOUGH, J. & TEICHMANN, S. A. 2001a. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology,* 310**,** 311-325.

APIC, G., GOUGH, J. & TEICHMANN, S. A. 2001b. An insight into domain combinations. *Bioinformatics,* 17**,** S83-S89.

APIC, G., HUBER, W. & TEICHMANN, S. A. 2003. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J Struct Funct Genomics,* 4**,** 67-78.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet,* 25.

AUYONG, A. S. M., FORD, R. & TAYLOR, P. W. J. 2015. The Role of Cutinase and its Impact on Pathogenicity of *Colletotrichum truncatum*. *Plant Pathology & Microbiology,* 6.

BALDWIN, T. K., WINNENBURG, R., URBAN, M., RAWLINGS, C., KOEHLER, J. & HAMMOND-KOSACK, K. E. 2006. The Pathogen-Host Interactions Database (PHI-base) Provides Insights into Generic and Novel Themes of Pathogenicity. *Molecular Plant-Microbe Interactions,* 19**,** 1451-1462.

BARABASI, A. L. & ALBERT, R. 1999. Emergence of scaling in random networks. *Science,* 286**,** 509-12.

BARKER, W. C., GEORGE, D. G., MEWES, H. W., PFEIFFER, F. & TSUGITA, A. 1993. The PIR-International databases. *Nucleic Acids Res,* 21**,** 3089-92.

BARMAN, R. K., SAHA, S. & DAS, S. 2014. Prediction of Interactions between Viral and Host Proteins Using Supervised Machine Learning Methods. *PLoS ONE,* 9**,** e112034.

BASU, M., CARMEL, L., ROGOZIN, I. & KOONIN, E. 2008. Evolution of protein domain promiscuity in eukaryotes. *Genome Research,* 18**,** 449 - 461.

BATEMAN, A., BIRNEY, E., CERRUTI, L., DURBIN, R., ETWILLER, L., EDDY, SEAN R., GRIFFITHS-JONES, S., HOWE, K. L., MARSHALL, M. & SONNHAMMER, E. L. L. 2002. The Pfam Protein Families Database. *Nucleic Acids Research,* 30**,** 276-280.

BATEMAN, A., COGGILL, P. & FINN, R. D. 2010. DUFs: families in search of function. *Acta Crystallographica Section F,* 66**,** 1148-1152.

BEN-DANIEL, B. H., BAR-ZVI, D. & TSROR LAHKIM, L. 2012. Pectate lyase affects pathogenicity in natural isolates of *Colletotrichum coccodes* and in pelA gene-disrupted and gene-overexpressing mutant lines. *Mol Plant Pathol,* 13**,** 187-97.

BENNETT, J. W. & KLICH, M. 2003. Mycotoxins. *Clinical Microbiology Reviews,* 16**,** 497-516.

BENNETT, L., KITTAS, A., MUIRHEAD, G., PAPAGEORGIOU, L. G. & TSOKA, S. 2015. Detection of composite communities in multiplex biological networks. *Sci Rep,* 5**,** 10345.

BENNETT, L., LYSENKO, A., PAPAGEORGIOU, L. G., URBAN, M., HAMMOND-KOSACK, K., RAWLINGS, C., SAQI, M. & TSOKA, S. 2012. Detection of Multi-clustered Genes and Community Structure for the Plant Pathogenic Fungus *Fusarium graminearum. In:* GILBERT, D. & HEINER, M. (eds.) *Computational Methods in Systems Biology: 10th International Conference, CMSB 2012, London, UK, October 3-5, 2012. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

BENT, A. F. & MACKEY, D. 2007. Elicitors, Effectors, and R Genes: The New Paradigm and a Lifetime Supply of Questions. *Annual Review of Phytopathology,* 45**,** 399-436.

BOCCALETTI, S., LATORA, V., MORENO, Y., CHAVEZ, M. & HWANG, D. U. 2006. Complex networks: Structure and dynamics. *Physics Reports,* 424**,** 175-308.

BOLIVAR, J. C., MACHENS, F., BRILL, Y., ROMANOV, A., BULOW, L. & HEHL, R. 2014. 'In silico expression analysis', a novel PathoPlant web tool to identify abiotic and biotic stress conditions associated with specific cis-regulatory sequences. *Database (Oxford),* 2014**,** bau030.

BORK, P. 1991. Shuffled domains in extracellular proteins. *FEBS Letters,* 286**,** 47-54.

BREIMAN, L. 2001. Random Forests. *Machine Learning,* 45**,** 5-32.

BROWN, N. A., ANTONIW, J. & HAMMOND-KOSACK, K. E. 2012a. The Predicted Secretome of the Plant Pathogenic Fungus *Fusarium graminearum*: A Refined Comparative Analysis. *PLoS ONE,* 7**,** e33731.

BROWN, N. A., URBAN, M. & AND HAMMOND-KOSACK, K. E. 2015. The trans-kingdom identification of negative regulators of pathogen hypervirulence. *FEMS Microbiological Letters (in press).*

BROWN, S. P., CORNFORTH, D. M. & MIDEO, N. 2012b. Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. *Trends in Microbiology,* 20**,** 336-342.

BUCHANAN M., CALDARELLI G., DE LOS RIOS P., RAO F. & M., V. 2010. *Networks in Cell Biology*, Cambridge University Press.

BULJAN, M. & BATEMAN, A. 2009. The evolution of protein domain families. *Biochem Soc Trans,* 37**,** 751-5.

CANTU, D., GOVINDARAJULU, M., KOZIK, A., WANG, M., CHEN, X., KOJIMA, K. K., JURKA, J., MICHELMORE, R. W. & DUBCOVSKY, J. 2011. Next Generation Sequencing Provides Rapid Access to the Genome of *Puccinia striiformis* f. sp. *tritici*, the Causal Agent of Wheat Stripe Rust. *PLoS ONE,* 6**,** e24230.

CATANZARITI, A. M., DODDS, P. N., LAWRENCE, G. J., AYLIFFE, M. A. & ELLIS, J. G. 2006. Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell,* 18**,** 243-56.

CHAISSON, M. J., BRINZA, D. & PEVZNER, P. A. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research,* 19**,** 336-346.

CHATR-ARYAMONTRI, A., BREITKREUTZ, B.-J., OUGHTRED, R., BOUCHER, L., HEINICKE, S., CHEN, D., STARK, C., BREITKREUTZ, A., KOLAS, N., O'DONNELL, L., REGULY, T., NIXON, J., RAMAGE, L., WINTER, A., SELLAM, A., CHANG, C., HIRSCHMAN, J., THEESFELD, C., RUST, J., LIVSTONE, M. S., DOLINSKI, K. & TYERS, M. 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Research,* 43**,** D470-D478.

CHEN, D., WANG, Y., ZHOU, X., WANG, Y. & XU, J. R. 2014. The Sch9 kinase regulates conidium size, stress responses, and pathogenesis in *Fusarium graminearum*. *PLoS One,* 9**,** e105811.

CHISHOLM, S. T., COAKER, G., DAY, B. & STASKAWICZ, B. J. 2006. Host-Microbe Interactions: Shaping the Evolution of the Plant Immune Response. *Cell,* 124**,** 803-814.

CHOI, S., JUN, H., BANG, J., CHUNG, S.-H., KIM, Y., KIM, B.-S., KIM, H., BEUCHAT, L. R. & RYU, J.-H. 2015. Behaviour of *Aspergillus flavus* and *Fusarium graminearum* on rice as affected by degree of milling, temperature, and relative humidity during storage. *Food Microbiology,* 46**,** 307-313.

CHOTHIA, C. 1992. One thousand families for the molecular biologist. *Nature,* 357**,** 543-544.

CHOTHIA, C., GOUGH, J., VOGEL, C. & TEICHMANN, S. 2003. Evolution of the protein repertoire. *Science,* 300**,** 1701 - 1703.

CHUMLEY, F. G. & VALENT, B. 1990. Genetic-Analysis of Melanin-Deficient, Nonpathogenic Mutants of *Magnaporthe-grisea*. *Molecular Plant-Microbe Interactions,* 3**,** 135-143.

CISSÉ, O. H., ALMEIDA, J. M. G. C. F., FONSECA, Á., KUMAR, A. A., SALOJÄRVI, J., OVERMYER, K., HAUSER, P. M. & PAGNI, M. 2013. Genome Sequencing of the Plant Pathogen *Taphrina deformans*, the Causal Agent of Peach Leaf Curl. *mBio,* 4.

CNOSSEN-FASSONI, A., BAZZOLLI, D. M., BROMMONSCHENKEL, S. H., FERNANDES DE ARAUJO, E. & DE QUEIROZ, M. V. 2013. The pectate lyase encoded by the pecCl1 gene is an important determinant for the aggressiveness of *Colletotrichum lindemuthianum*. *J Microbiol,* 51**,** 461-70.

COHEN R., H. S. 2010. *Complex Networks: Structure, Robustness and Function*, Cambridge University Press.

CONSORTIUM, T. U. 2014. UniProt: a hub for protein information. *Nucleic Acids Research*.

CONSORTIUM, T. U. 2015. UniProt: a hub for protein information. *Nucleic Acids Research,* 43**,** D204-D212.

COPLEY, R. R., DOERKS, T., LETUNIC, I. & BORK, P. 2002. Protein domain analysis in the era of complete genomes. *FEBS Letters,* 513**,** 129-134.

CUOMO, C. A., GULDENER, U., XU, J. R., TRAIL, F., TURGEON, B. G., DI PIETRO, A., WALTON, J. D., MA, L. J., BAKER, S. E., REP, M., ADAM, G., ANTONIW, J., BALDWIN, T., CALVO, S., CHANG, Y. L., DECAPRIO, D., GALE, L. R., GNERRE, S., GOSWAMI, R. S., HAMMOND-KOSACK, K., HARRIS, L. J., HILBURN, K., KENNELL, J. C., KROKEN, S., MAGNUSON, J. K., MANNHAUPT, G., MAUCELI, E., MEWES, H. W., MITTERBAUER, R., MUEHLBAUER, G., MUNSTERKOTTER, M., NELSON, D., O'DONNELL, K., OUELLET, T., QI, W., QUESNEVILLE, H., RONCERO, M. I., SEONG, K. Y., TETKO, I. V., URBAN, M., WAALWIJK, C., WARD, T. J., YAO, J., BIRREN, B. W. & KISTLER, H. C. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science,* 317**,** 1400-2.

CZEMBOR, E., STĘPIEŃ, Ł. & WAŚKIEWICZ, A. 2015. Effect of environmental factors on *Fusarium* species and associated mycotoxins in maize grain grown in Poland. PLoS ONE, 10, e0133644.

DASH, S., VAN HEMERT, J., HONG, L., WISE, R. P. & DICKERSON, J. A. 2012. PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Research,* 40**,** D1194-D1201.

DAYHOFF, M. O. 1974. Computer analysis of protein sequences. *In:* SILER, W. & LINDBERG, D. B. (eds.) *Computers in Life Science Research.* Springer US.

DAYHOFF, M. O. 1976. The origin and evolution of protein superfamilies. *Federation proceedings,* 35**,** 2132-2138.

DAYHOFF, M. O., MCLAUGHLIN, P. J., BARKER, W. C. & HUNT, L. T. 1975. Evolution of sequences within protein superfamilies. *Naturwissenschaften,* 62**,** 154-161.

DE LUCCA, A. J. 2007. Harmful fungi in both agriculture and medicine. *Rev Iberoam Micol,* 24**,** 3-13.

DE SILVA, E. & STUMPF, M. P. H. 2005. Complex networks and simple models in biology. *Journal of the Royal Society Interface,* 2**,** 419-430.

DEAN, R., TALBOT, N., EBBOLE, D., FARMAN, M., MITCHELL, T., ORBACH, M., THON, M., KULKARNI, R., XU, J., PAN, H., READ, N., LEE, Y., CARBONE, I., BROWN, D., OH, Y., DONOFRIO, N., JEONG, J., SOANES, D., DJONOVIC, S., KOLOMIETS, E., REHMEYER, C., LI, W., HARDING, M., KIM, S., LEBRUN, M., BOHNERT, H., COUGHLAN, S., BUTLER, J., CALVO, S., MA, L., NICOL, R., PURCELL, S., NUSBAUM, C., GALAGAN, J. & BIRREN, B. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature,* 434**,** 980 - 986.

DEISING, H. B., WERNER, S. & WERNITZ, M. 2000. The role of fungal appressoriain plant infection. *Microbes and Infection,* 2**,** 1631-1641.

DENISOV, G., WALENZ, B., HALPERN, A. L., MILLER, J., AXELROD, N., LEVY, S. & SUTTON, G. 2008. Consensus generation and variant detection by Celera Assembler. *Bioinformatics,* 24**,** 1035-40.

DIGUISTINI, S., LIAO, N., PLATT, D., ROBERTSON, G., SEIDEL, M., CHAN, S., DOCKING, T. R., BIROL, I., HOLT, R., HIRST, M., MARDIS, E., MARRA, M., HAMELIN, R., BOHLMANN, J., BREUIL, C. & JONES, S. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology,* 10**,** R94.

DJAMEI, A. & KAHMANN, R. 2012. *Ustilago maydis*: dissecting the molecular interface between pathogen and plant. *PLoS Pathog,* 8**,** e1002955.

DOHM, J. C., LOTTAZ, C., BORODINA, T. & HIMMELBAUER, H. 2007. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res,* 17**,** 1697-706.

DONG, S., STAM, R., CANO, L. M., SONG, J., SKLENAR, J., YOSHIDA, K., BOZKURT, T. O., OLIVA, R., LIU, Z., TIAN, M., WIN, J., BANFIELD, M. J., JONES, A. M. E., VAN DER HOORN, R. A. L. & KAMOUN, S. 2014. Effector Specialization in a Lineage of the Irish Potato Famine Pathogen. *Science,* 343**,** 552-555.

DOSS, R. P. 1999. Composition and Enzymatic Activity of the Extracellular Matrix Secreted by Germlings of *Botrytis cinerea*. *Applied and Environmental Microbiology,* 65**,** 404-408.

DURMUŞ, S., ÇAKıR, T., ÖZGÜR, A. & GUTHKE, R. 2015. A review on computational systems biology of pathogen–host interactions. *Frontiers in Microbiology,* 6**,** 235.

DVORACKOVA, I. & PICHOVA, V. 1986. Pulmonary interstitial fibrosis with evidence of aflatoxin B1 in lung tissue. *Journal of Toxicology and Environmental Health,* 1**,** 153-157.

DYER, M., MURALI, T. & SOBRAL, B. 2007. Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics,* 23**,** i159 - 166.

EBBOLE, D. 2007. *Magnaporthe* as a model for understanding host-pathogen interactions. *Annu Rev Phytopathol,* 45**,** 437 - 456.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

EKMAN, D., BJORKLUND, A. K., FREY-SKOTT, J. & ELOFSSON, A. 2005. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol,* 348**,** 231-43.

ENRIGHT, A. J., VAN DONGEN, S. & OUZOUNIS, C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res,* 30**,** 1575-84.

FAWKE, S., DOUMANE, M. & SCHORNACK, S. 2015. Oomycete Interactions with Plants: Infection Strategies and Resistance Principles. *Microbiology and Molecular Biology Reviews,* 79**,** 263-280.

FINN, R. D., BATEMAN, A., CLEMENTS, J., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., HEGER, A., HETHERINGTON, K., HOLM, L., MISTRY, J., SONNHAMMER, E. L. L., TATE, J. & PUNTA, M. 2014a. Pfam: the protein families database. *Nucleic Acids Research,* 42**,** D222-D230.

FINN, R. D., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., MISTRY, J., MITCHELL, A. L., POTTER, S. C., PUNTA, M., QURESHI, M., SANGRADOR-VEGAS, A., SALAZAR, G. A., TATE, J. & BATEMAN, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research,* 44**,** D279-D285.

FINN, R. D., MILLER, B. L., CLEMENTS, J. & BATEMAN, A. 2014b. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research,* 42**,** D364-D373.

FISHER, M. C., HENK, D. A., BRIGGS, C. J., BROWNSTEIN, J. S., MADOFF, L. C., MCCRAW, S. L. & GURR, S. J. 2012. Emerging fungal threats to animal, plant and ecosystem health. *Nature,* 484**,** 186-194.

FORTUNATO, S. 2010. Community detection in graphs. *Physics Reports,* 486**,** 75-174.

FORTUNATO, S. & CASTELLANO, C. 2012. Community Structure in Graphs. *In:* MEYERS, R. A. (ed.) *Computational Complexity.* Springer New York.

GAFFOOR, I., BROWN, D. W., PLATTNER, R., PROCTOR, R. H., QI, W. & TRAIL, F. 2005. Functional analysis of the polyketide synthase genes in the filamentous fungus *Gibberella zeae* (anamorph *Fusarium graminearum*). *Eukaryot Cell,* 4**,** 1926-33.

GALAGAN, J. E., HENN, M. R., MA, L. J., CUOMO, C. A. & BIRREN, B. 2005. Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res,* 15**,** 1620-31.

GALAGAN, J. E. & SELKER, E. U. 2004. RIP: the evolutionary cost of genome defense. *Trends in Genetics,* 20**,** 417-423.

GARDINER, D. M., KAZAN, K. & MANNERS, J. M. 2009. Novel genes of *Fusarium graminearum* that negatively regulate deoxynivalenol production and virulence. *Mol Plant Microbe Interact,* 22**,** 1588-600.

GARDINER, S. A., BODDU, J., BERTHILLER, F., HAMETNER, C., STUPAR, R. M., ADAM, G. & MUEHLBAUER, G. J. 2010. Transcriptome analysis of the barley-deoxynivalenol interaction: evidence for a role of glutathione in deoxynivalenol detoxification. *Mol Plant Microbe Interact,* 23**,** 962-76.

GEISLER-LEE, J., O'TOOLE, N., AMMAR, R., PROVART, N. J., MILLAR, A. H. & GEISLER, M. 2007. A Predicted Interactome for *Arabidopsis*. *Plant Physiology,* 145**,** 317-329.

GHOORAH, A. W., DEVIGNES, M. D., SMAIL-TABBONE, M. & RITCHIE, D. W. 2014. KBDOCK 2013: a spatial classification of 3D protein domain family interactions. *Nucleic Acids Res,* 42**,** D389-95.

GILLIS, J. & PAVLIDIS, P. 2012. “Guilt by Association” Is the Exception Rather Than the Rule in Gene Networks. *PLOS Computational Biology,* 8**,** e1002444.

GIRVAN, M. & NEWMAN, M. E. J. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences,* 99**,** 7821-7826.

GOFFEAU, A., BARRELL, B. G., BUSSEY, H., DAVIS, R. W., DUJON, B., FELDMANN, H., GALIBERT, F., HOHEISEL, J. D., JACQ, C., JOHNSTON, M., LOUIS, E. J., MEWES, H. W., MURAKAMI, Y., PHILIPPSEN, P., TETTELIN, H. & OLIVER, S. G. 1996. Life with 6000 genes. *Science,* 274**,** 546, 563-7.

GOMES, L. B., WARD, T. J., BADIALE-FURLONG, E. & DEL PONTE, E. M. 2015. Species composition, toxigenic potential and pathogenicity of *Fusarium graminearum* species complex isolates from southern Brazilian rice. *Plant Pathology,* 64**,** 980-987.

GONZÁLEZ, M., BRITO, N., FRÍAS, M. & GONZÁLEZ, C. 2013. Botrytis cinerea Protein O-Mannosyltransferases Play Critical Roles in Morphogenesis, Growth, and Virulence. *PLoS ONE,* 8**,** e65924.

GOODACRE, N. F., GERLOFF, D. L. & UETZ, P. 2014. Protein domains of unknown function are essential in bacteria. *MBio,* 5**,** e00744-13.

GOSWAMI, R. S. & KISTLER, H. C. 2004. Heading for disaster: *Fusarium graminearum* on cereal crops. *Molecular Plant Pathology,* 5**,** 515-525.

GOUGH, J. 2006. Genomic scale sub-family assignment of protein domains. *Nucleic Acids Res,* 34**,** 3625-33.

GRIGORIEV, I. V., CULLEN, D., GOODWIN, S. B., HIBBETT, D., JEFFRIES, T. W., KUBICEK, C. P., KUSKE, C., MAGNUSON, J. K., MARTIN, F., SPATAFORA, J. W., TSANG, A. & BAKER, S. E. 2011. Fueling the future with fungal genomics. *Mycology,* 2**,** 192-209.

GU, H., ZHU, P., JIAO, Y., MENG, Y. & CHEN, M. 2011. PRIN: a predicted rice interactome network. *BMC Bioinformatics,* 12**,** 161.

GUIMERA, R. & NUNES AMARAL, L. A. 2005. Functional cartography of complex metabolic networks. *Nature,* 433**,** 895-900.

GULDENER, U., MANNHAUPT, G., MUNSTERKOTTER, M., HAASE, D., OESTERHELD, M., STUMPFLEN, V., MEWES, H. W. & ADAM, G. 2006a. FGDB: a comprehensive fungal genome resource on the plant pathogen *Fusarium graminearum*. *Nucleic Acids Res,* 34**,** D456-8.

GULDENER, U., SEONG, K. Y., BODDU, J., CHO, S., TRAIL, F., XU, J. R., ADAM, G., MEWES, H. W., MUEHLBAUER, G. J. & KISTLER, H. C. 2006b. Development of a *Fusarium graminearum* Affymetrix GeneChip for profiling fungal gene expression in vitro and in planta. *Fungal Genet Biol,* 43**,** 316-25.

GUO, L., ZHAO, G., XU, J. R., KISTLER, H. C., GAO, L. & MA, L. J. 2016. Compartmentalized gene regulatory network of the pathogenic fungus *Fusarium graminearum*. *New Phytol,* 211**,** 527-41.

GUPTA, V. K., CHATTOPADHYAY, P., KALITA, M. C., CHAURASIA, A. K., GOGOI, H. K. & SINGH, L. 2011. Isolation and determination of deoxynivalenol by reversed-phase high-pressure liquid chromatography. *Pharmaceutical Methods,* 2**,** 25-29.

GUTLEB, A. C., MORRISON, E. & MURK, A. J. 2002. Cytotoxicity assays for mycotoxins produced by *Fusarium* strains: a review. *Environmental Toxicology and Pharmacology,* 11**,** 309-320.

HAMED, M. A. & ALI, S. A. 2013. Non-viral factors contributing to hepatocellular carcinoma. *World Journal of Hepatology,* 5**,** 311-322.

HANSEN, F. T., SØRENSEN, J. L., GIESE, H., SONDERGAARD, T. E. & FRANDSEN, R. J. N. 2012. Quick guide to polyketide synthase and nonribosomal synthetase genes in *Fusarium*. *International Journal of Food Microbiology,* 155**,** 128-136.

HE, F., ZHANG, Y., CHEN, H., ZHANG, Z. & PENG, Y.-L. 2008. The prediction of protein-protein interaction networks in rice blast fungus. *BMC Genomics,* 9**,** 519.

HEGYI, H. & GERSTEIN, M. 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol,* 288**,** 147-64.

HERRERO, J., MUFFATO, M., BEAL, K., FITZGERALD, S., GORDON, L., PIGNATELLI, M., VILELLA, A. J., SEARLE, S. M. J., AMODE, R., BRENT, S., SPOONER, W., KULESHA, E., YATES, A. & FLICEK, P. 2016. Ensembl comparative genomics resources. *Database,* 2016**,** bav096-bav096.

HIFNAWY, M. S., MANGOUD, A. M., EISSA, M. H., NOR EDIN, E., MOSTAFA, Y., ABOUEL-MAGD, Y., SABEE, E. I., AMIN, I., ISMAIL, A., MORSY, T. A., MAHROUS, S., AFEFY, A. F., EL-SHORBAGY, E., EL-SADAWY, M., RAGAB, H., HASSAN, M. I., EL-HADY, G. & SABER, M. 2004. The role of aflatoxin-contaminated food materials and HCV in developing hepatocellular carcinoma in Al-Sharkia Governorate, Egypt. *J Egypt Soc Parasitol,* 34**,** 479-88.

HOLZ, J. F. F. G. 1995. Initial Infection Processes by *Botrytis cinerea* on Nectarine and Plum Fruit and the Development of Decay

Phytopathology, 85**,** 82-87.

HORTON, P., PARK, K., OBAYASHI, T., FUJITA, N., HARADA, H., ADAMS-COLLIER, C. & NAKAI, K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res***,** W585 - 587.

HOWARD, R. J., FERRARI, M. A., ROACH, D. H. & MONEY, N. P. 1991. Penetration of hard substrates by a fungus employing enormous turgor pressures. *Proceedings of the National Academy of Sciences,* 88**,** 11281-11284.

HOWARD, R. J. & VALENT, B. 1996. BREAKING AND ENTERING: Host Penetration by the Fungal Rice Blast Pathogen *Magnaporthe grisea*. *Annual Review of Microbiology,* 50**,** 491-512.

HUEZA, I., RASPANTINI, P., RASPANTINI, L., LATORRE, A. & GÓRNIAK, S. 2014. Zearalenone, an Estrogenic Mycotoxin, Is an Immunotoxic Compound. *Toxins,* 6**,** 1080.

HUSSEIN, H. S. & BRASEL, J. M. 2001. Toxicity, metabolism, and impact of mycotoxins on humans and animals. *Toxicology,* 167**,** 101-134.

JANSEN, R., YU, H., GREENBAUM, D., KLUGER, Y., KROGAN, N. J., CHUNG, S., EMILI, A., SNYDER, M., GREENBLATT, J. F. & GERSTEIN, M. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science,* 302**,** 449-53.

JAROSZEWSKI, L., LI, Z., KRISHNA, S. S., BAKOLITSA, C., WOOLEY, J., DEACON, A. M., WILSON, I. A. & GODZIK, A. 2009. Exploration of Uncharted Regions of the Protein Universe. *PLoS Biol,* 7**,** e1000205.

JECK, W. R., REINHARDT, J. A., BALTRUS, D. A., HICKENBOTHAM, M. T., MAGRINI, V., MARDIS, E. R., DANGL, J. L. & JONES, C. D. 2007. Extending assembly of short DNA sequences to handle error. *Bioinformatics,* 23**,** 2942-4.

JENCZMIONKA, N. J., MAIER, F. J., LOSCH, A. P. & SCHAFER, W. 2003. Mating, conidiation and pathogenicity of *Fusarium graminearum*, the main causal agent of the head-blight disease of wheat, are regulated by the MAP kinase gpmk1. *Curr Genet,* 43**,** 87-95.

JEONG, H., MASON, S. P., BARABASI, A. L. & OLTVAI, Z. N. 2001. Lethality and centrality in protein networks. *Nature,* 411**,** 41-42.

JERMY, A. 2012. Fungal pathogenesis: *Magnaporthe* puts a ring on it. *Nat Rev Micro,* 10**,** 521-521.

JONES, J. D. G. & DANGL, J. L. 2006. The plant immune system. *Nature,* 444**,** 323-329.

KERSEY, P. J., ALLEN, J. E., ALLOT, A., BARBA, M., BODDU, S., BOLT, B. J., CARVALHO-SILVA, D., CHRISTENSEN, M., DAVIS, P., GRABMUELLER, C., KUMAR, N., LIU, Z., MAUREL, T., MOORE, B., MCDOWALL, M. D., MAHESWARI, U., NAAMATI, G., NEWMAN, V., ONG, C. K., PAULINI, M., PEDRO, H., PERRY, E., RUSSELL, M., SPARROW, H., TAPANARI, E., TAYLOR, K., VULLO, A., WILLIAMS, G., ZADISSIA, A., OLSON, A., STEIN, J., WEI, S., TELLO-RUIZ, M., WARE, D., LUCIANI, A., POTTER, S., FINN, R. D., URBAN, M., HAMMOND-KOSACK, K. E., BOLSER, D. M., DE SILVA, N., HOWE, K. L., LANGRIDGE, N., MASLEN, G., STAINES, D. M. & YATES, A. 2018. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res,* 46**,** D802-d808.

KERSEY, P. J., ALLEN, J. E., CHRISTENSEN, M., DAVIS, P., FALIN, L. J., GRABMUELLER, C., HUGHES, D. S., HUMPHREY, J., KERHORNOU, A., KHOBOVA, J., LANGRIDGE, N., MCDOWALL, M. D., MAHESWARI, U., MASLEN, G., NUHN, M., ONG, C. K., PAULINI, M., PEDRO, H., TONEVA, I., TULI, M. A., WALTS, B., WILLIAMS, G., WILSON, D., YOUENS-CLARK, K., MONACO, M. K., STEIN, J., WEI, X., WARE, D., BOLSER, D. M., HOWE, K. L., KULESHA, E., LAWSON, D. & STAINES, D. M. 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res,* 42**,** D546-52.

KESHAVA PRASAD, T. S., GOEL, R., KANDASAMY, K., KEERTHIKUMAR, S., KUMAR, S., MATHIVANAN, S., TELIKICHERLA, D., RAJU, R., SHAFREEN, B., VENUGOPAL, A., BALAKRISHNAN, L., MARIMUTHU, A., BANERJEE, S., SOMANATHAN, D. S., SEBASTIAN, A., RANI, S., RAY, S., HARRYS KISHORE, C. J., KANTH, S., AHMED, M., KASHYAP, M. K., MOHMOOD, R., RAMACHANDRA, Y. L., KRISHNA, V., RAHIMAN, B. A., MOHAN, S., RANGANATHAN, P., RAMABADRAN, S., CHAERKADY, R. & PANDEY, A. 2009. Human Protein Reference Database--2009 update. *Nucleic Acids Res,* 37**,** D767-72.

KEW, M. C. 2013. Aflatoxins as a cause of hepatocellular carcinoma. *J Gastrointestin Liver Dis,* 22**,** 305-10.

KIM, J.-E., MYONG, K., SHIM, W.-B., YUN, S.-H. & LEE, Y.-W. 2007. Functional characterization of acetylglutamate synthase and phosphoribosylamine-glycine ligase genes in *Gibberella zeae*. *Current Genetics,* 51**,** 99-108.

KIM, Y. T., LEE, Y. R., JIN, J., HAN, K. H., KIM, H., KIM, J. C., LEE, T., YUN, S. H. & LEE, Y. W. 2005. Two different polyketide synthase genes are required for synthesis of zearalenone in *Gibberella zeae*. *Mol Microbiol,* 58**,** 1102-13.

KING, R., URBAN, M., HAMMOND-KOSACK, M. C. U., HASSANI-PAK, K. & HAMMOND-KOSACK, K. E. 2015. The completed genome sequence of the pathogenic ascomycete fungus *Fusarium graminearum*. *BMC Genomics,* 16**,** 544.

KNOGGE, W. 1996. Fungal Infection of Plants. *The Plant Cell,* 8**,** 1711-1722.

KOECK, M., HARDHAM, A. R. & DODDS, P. N. 2011. The role of effectors of biotrophic and hemibiotrophic fungi in infection. *Cellular microbiology,* 13**,** 1849-1857.

KÖHLER, J., BAUMBACH, J., TAUBERT, J., SPECHT, M., SKUSA, A., RÜEGG, A., RAWLINGS, C., VERRIER, P. & PHILIPPI, S. 2006. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics,* 22**,** 1383-1390.

KÖHLER, S., BAUER, S., HORN, D. & ROBINSON, P. N. 2008. Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics,* 82**,** 949-958.

KONC, J., HODOŠČEK, M., OGRIZEK, M., TRYKOWSKA KONC, J. & JANEŽIČ, D. 2013. Structure-Based Function Prediction of Uncharacterized Protein Using Binding Sites Comparison. *PLoS Comput Biol,* 9**,** e1003341.

KOONIN, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet,* 39**,** 309-38.

KOTCHONI, S. O., JIMENEZ-LOPEZ, J. C., GAO, D., EDWARDS, V., GACHOMO, E. W., MARGAM, V. M. & SEUFFERHELD, M. J. 2010. Modeling-Dependent Protein Characterization of the Rice Aldehyde Dehydrogenase (ALDH) Superfamily Reveals Distinct Functional and Structural Features. *PLoS ONE,* 5**,** e11516.

KROGH, A., BROWN, M., MIAN, I. S., SJOLANDER, K. & HAUSSLER, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol,* 235**,** 1501-31.

LAM, S. D., DAWSON, N. L., DAS, S., SILLITOE, I., ASHFORD, P., LEE, D., LEHTINEN, S., ORENGO, C. A. & LEES, J. G. 2016. Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res,* 44**,** D404-9.

LEE, J., CHANG, I.-Y., KIM, H., YUN, S.-H., LESLIE, J. F. & LEE, Y.-W. 2009. Genetic Diversity and Fitness of *Fusarium graminearum* Populations from Rice in Korea. *Applied and Environmental Microbiology,* 75**,** 3289-3295.

LEE, S., CHAN, C., TSAI, C., LAI, J., WANG, F., KAO, C. & HUANG, C. 2008. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics,* 12**,** S11.

LENG, Y. & ZHONG, S. 2015. The Role of Mitogen-Activated Protein (MAP) Kinase Signaling Components in the Fungal Development, Stress Response and Virulence of the Fungal Cereal Pathogen *Bipolaris sorokiniana*. *PLoS One,* 10**,** e0128291.

LIANG, S., ZHENG, D., STANDLEY, D., GUO, H. & ZHANG, C. 2013. A novel function prediction approach using protein overlap networks. *BMC Systems Biology,* 7**,** 61.

LICATA, L., BRIGANTI, L., PELUSO, D., PERFETTO, L., IANNUCCELLI, M., GALEOTA, E., SACCO, F., PALMA, A., NARDOZZA, A. P., SANTONICO, E., CASTAGNOLI, L. & CESARENI, G. 2012. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res,* 40**,** D857-61.

LIU, X., TANG, W.-H., ZHAO, X.-M. & CHEN, L. 2010. A Network Approach to Predict Pathogenic Genes for *Fusarium graminearum. PLoS ONE,* 5**,** e13021.

LO CONTE, L., AILEY, B., HUBBARD, T. J. P., BRENNER, S. E., MURZIN, A. G. & CHOTHIA, C. 2000. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Research,* 28**,** 257-259.

LUDWIG, N., LÖHRER, M., HEMPEL, M., MATHEA, S., SCHLIEBNER, I., MENZEL, M., KIESOW, A., SCHAFFRATH, U., DEISING, H. B. & HORBACH, R. 2013. Melanin Is Not Required for Turgor Generation but Enhances Cell-Wall Rigidity in Appressoria of the Corn Pathogen *Colletotrichum graminicola*. *Molecular Plant-Microbe Interactions,* 27**,** 315-327.

LYSENKO, A., DEFOIN-PLATEL, M., HASSANI-PAK, K., TAUBERT, J., HODGMAN, C., RAWLINGS, C. & SAQI, M. 2011. Assessing the functional coherence of modules found in multiple-evidence networks from *Arabidopsis*. *BMC Bioinformatics,* 12**,** 203.

LYSENKO, A., HINDLE, M., TAUBERT, J., SAQI, M. & RAWLINGS, C. 2009. Data integration for plant genomics--exemplars from the integration of *Arabidopsis thaliana* databases. *Briefings in Bioinformatics,* 10**,** 676 - 693.

LYSENKO, A., URBAN, M., BENNETT, L., TSOKA, S., JANOWSKA-SEJDA, E., RAWLINGS, C. J., HAMMOND-KOSACK, K. E. & SAQI, M. 2013. Network-Based Data Integration for Selecting Candidate Virulence Associated Proteins in the Cereal Infecting Fungus *Fusarium graminearum*. *PLoS ONE,* 8**,** e67926.

LYSOE, E., KLEMSDAL, S. S., BONE, K. R., FRANDSEN, R. J., JOHANSEN, T., THRANE, U. & GIESE, H. 2006. The PKS4 gene of *Fusarium graminearum* is essential for zearalenone production. *Appl Environ Microbiol,* 72**,** 3924-32.

MA, H. & ZENG, A. P. 2003. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics,* 19**,** 270-7.

MA, L.-J., VAN DER DOES, H. C., BORKOVICH, K. A., COLEMAN, J. J., DABOUSSI, M.-J., DI PIETRO, A., DUFRESNE, M., FREITAG, M., GRABHERR, M., HENRISSAT, B., HOUTERMAN, P. M., KANG, S., SHIM, W.-B., WOLOSHUK, C., XIE, X., XU, J.-R., ANTONIW, J., BAKER, S. E., BLUHM, B. H., BREAKSPEAR, A., BROWN, D. W., BUTCHKO, R. A. E., CHAPMAN, S., COULSON, R., COUTINHO, P. M., DANCHIN, E. G. J., DIENER, A., GALE, L. R., GARDINER, D. M., GOFF, S., HAMMOND-KOSACK, K. E., HILBURN, K., HUA-VAN, A., JONKERS, W., KAZAN, K., KODIRA, C. D., KOEHRSEN, M., KUMAR, L., LEE, Y.-H., LI, L., MANNERS, J. M., MIRANDA-SAAVEDRA, D., MUKHERJEE, M., PARK, G., PARK, J., PARK, S.-Y., PROCTOR, R. H., REGEV, A., RUIZ-ROLDAN, M. C., SAIN, D., SAKTHIKUMAR, S., SYKES, S., SCHWARTZ, D. C., TURGEON, B. G., WAPINSKI, I., YODER, O., YOUNG, S., ZENG, Q., ZHOU, S., GALAGAN, J., CUOMO, C. A., KISTLER, H. C. & REP, M. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature,* 464**,** 367-373.

MANNING CAS, H. 1999. Foundations of Statistical Natural Language Processing. *Cambridge, MA: MIT Press*.

MARASAS, W. F. O., RILEY, R. T., HENDRICKS, K. A., STEVENS, V. L., SADLER, T. W., GELINEAU-VAN WAES, J., MISSMER, S. A., CABRERA, J., TORRES, O., GELDERBLOM, W. C. A., ALLEGOOD, J., MARTINEZ, C., MADDOX, J., MILLER, J. D., STARR, L., SULLARDS, M. C., ROMAN, A. V., VOSS, K. A., WANG, E. & MERRILL, A. H. 2004. Fumonisins disrupt sphingolipid metabolism, folate transport, and neural tube development in embryo culture and in vivo: A potential risk factor for human neural tube defects among populations consuming fumonisin-contaminated maize. *Journal of Nutrition,* 134**,** 711-716.

MARIN, S., RAMOS, A. J., CANO-SANCHO, G. & SANCHIS, V. 2013. Mycotoxins: Occurrence, toxicology, and exposure assessment. *Food and Chemical Toxicology,* 60**,** 218-237.

MARSHALL, R., KOMBRINK, A., MOTTERAM, J., LOZA-REYES, E., LUCAS, J., HAMMOND-KOSACK, K. E., THOMMA, B. P. H. J. & RUDD, J. J. 2011. Analysis of Two in Planta Expressed LysM Effector Homologs from the Fungus *Mycosphaerella graminicola* Reveals Novel Functional Properties and Varying Contributions to Virulence on Wheat. *Plant Physiology,* 156**,** 756-769.

MATTHEWS, L., VAGLIO, P., REBOUL, J., GE, H., DAVIS, B., GARRELS, J., VINCENT, S. & VIDAL, M. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res,* 11**,** 2120 - 2126.

MEIER, J. L. & BURKART, M. D. 2009. The chemical biology of modular biosynthetic enzymes. *Chemical Society Reviews,* 38**,** 2012-2045.

MENTLAK, T. A., KOMBRINK, A., SHINYA, T., RYDER, L. S., OTOMO, I., SAITOH, H., TERAUCHI, R., NISHIZAWA, Y., SHIBUYA, N., THOMMA, B. P. H. J. & TALBOT, N. J. 2012. Effector-Mediated Suppression of Chitin-Triggered Immunity by *Magnaporthe oryzae* Is Necessary for Rice Blast Disease. *The Plant Cell,* 24**,** 322-335.

MERRILL, A. H., SULLARDS, M. C., WANG, E., VOSS, K. A. & RILEY, R. T. 2001. Sphingolipid metabolism: Roles in signal transduction and disruption by fumonisins. *Environmental Health Perspectives,* 109**,** 283-289.

MILLER, J. R., KOREN, S. & SUTTON, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics,* 95**,** 315-327.

MIYARA, I., SHAFRAN, H., KRAMER HAIMOVICH, H., ROLLINS, J., SHERMAN, A. & PRUSKY, D. 2008. Multi-factor regulation of pectate lyase secretion by *Colletotrichum gloeosporioides* pathogenic on avocado fruits. *Mol Plant Pathol,* 9**,** 281-91.

MOSCA, R., CÉOL, A., STEIN, A., OLIVELLA, R. & ALOY, P. 2014. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research,* 42**,** D374-D379.

MUDGAL, R., SANDHYA, S., CHANDRA, N. & SRINIVASAN, N. 2015. De-DUFing the DUFs: Deciphering distant evolutionary relationships of Domains of Unknown Function using sensitive homology detection methods. *Biol Direct,* 10**,** 38.

MUKHTAR, M. S., CARVUNIS, A. R., DREZE, M., EPPLE, P., STEINBRENNER, J., MOORE, J., TASAN, M., GALLI, M., HAO, T., NISHIMURA, M. T., PEVZNER, S. J., DONOVAN, S. E., GHAMSARI, L., SANTHANAM, B., ROMERO, V., POULIN, M. M., GEBREAB, F., GUTIERREZ, B. J., TAM, S., MONACHELLO, D., BOXEM, M., HARBORT, C. J., MCDONALD, N., GAI, L., CHEN, H., HE, Y., VANDENHAUTE, J., ROTH, F. P., HILL, D. E., ECKER, J. R., VIDAL, M., BEYNON, J., BRAUN, P. & DANGL, J. L. 2011. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science,* 333**,** 596-601.

MURZIN, A., BRENNER, S., HUBBARD, T. & CHOTHIA, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequaences and structures. *J Mol Biol,* 247**,** 536 - 540.

NEMATI, M., MEHRAN, M. A., HAMED, P. K. & MASOUD, A. 2010. A survey on the occurrence of aflatoxin M1 in milk samples in Ardabil, Iran. *Food Control,* 21**,** 1022-1024.

NGUYEN, T. P. & HO, T. B. 2008. An integrative domain-based approach to predicting protein-protein interactions. *J Bioinform Comput Biol,* 6**,** 1115-32.

OATES, M. E., STAHLHACKE, J., VAVOULIS, D. V., SMITHERS, B., RACKHAM, O. J. L., SARDAR, A. J., ZAUCHA, J., THURLBY, N., FANG, H. & GOUGH, J. 2015. The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Research,* 43**,** D227-D233.

OPSAHL, T., AGNEESSENS, F. & SKVORETZ, J. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks,* 32**,** 245-251.

ORCHARD, S., AMMARI, M., ARANDA, B., BREUZA, L., BRIGANTI, L., BROACKES-CARTER, F., CAMPBELL, N. H., CHAVALI, G., CHEN, C., DEL-TORO, N., DUESBURY, M., DUMOUSSEAU, M., GALEOTA, E., HINZ, U., IANNUCCELLI, M., JAGANNATHAN, S., JIMENEZ, R., KHADAKE, J., LAGREID, A., LICATA, L., LOVERING, R. C., MELDAL, B., MELIDONI, A. N., MILAGROS, M., PELUSO, D., PERFETTO, L., PORRAS, P., RAGHUNATH, A., RICARD-BLUM, S., ROECHERT, B., STUTZ, A., TOGNOLLI, M., VAN ROEY, K., CESARENI, G. & HERMJAKOB, H. 2014. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research,* 42**,** D358-63.

ORENGO, C. A., MICHIE, A. D., JONES, S., JONES, D. T., SWINDELLS, M. B. & THORNTON, J. M. 1997. CATH--a hierarchic classification of protein domain structures. *Structure,* 5**,** 1093-108.

PARK, A. R., CHO, A.-R., SEO, J.-A., MIN, K., SON, H., LEE, J., CHOI, G. J., KIM, J.-C. & LEE, Y.-W. 2012. Functional analyses of regulators of G protein signaling in *Gibberella zeae*. *Fungal Genetics and Biology,* 49**,** 511-520.

PATHIRANA, U. P. D., WIMALASIRI, K. M. S., SILVA, K. F. S. T. & GUNARATHNE, S. P. 2010. Investigation of farm gate cow milk for aflatoxin M1. *Tropical Agricultural Research,* 21**,** 119-125.

PEDRO, H., MAHESWARI, U., URBAN, M., IRVINE, A. G., ALAYNE CUZICK, A., MCDOWALL, M. D., STAINES, M. D., KULESHA, E., HAMMOND-KOSACK, K. E. & AND KERSEY, P. J. 2016. PhytoPath: an integrative resource for plant pathogen genomics. *Nucleic Acids Research (accepted)*

PENN, T. J., WOOD, M. E., SOANES, D. M., CSUKAI, M., CORRAN, A. J. & TALBOT, N. J. 2015. Protein kinase C is essential for viability of the rice blast fungus *Magnaporthe oryzae*. *Molecular Microbiology***,** n/a-n/a.

PISTOL, G. C., BRAICU, C., MOTIU, M., GRAS, M. A., MARIN, D. E., STANCU, M., CALIN, L., ISRAEL-ROMING, F., BERINDAN-NEAGOE, I. & TARANU, I. 2015. Zearalenone Mycotoxin Affects Immune Mediators, MAPK Signalling Molecules, Nuclear Receptors and Genome-Wide Gene Expression in Pig Spleen. *PLoS ONE,* 10**,** e0127503.

POSTEL, S. & KEMMERLING, B. 2009. Plant systems for recognition of pathogen-associated molecular patterns. *Seminars in Cell & Developmental Biology,* 20**,** 1025-1031.

PROCTOR, R. H., HOHN, T. M. & MCCORMICK, S. P. 1995. Reduced virulence of *Gibberella zeae* caused by disruption of a trichothecene toxin biosynthetic gene. *Mol Plant Microbe Interact,* 8**,** 593-601.

PUNTA, M., COGGILL, P. C., EBERHARDT, R. Y., MISTRY, J., TATE, J., BOURSNELL, C., PANG, N., FORSLUND, K., CERIC, G., CLEMENTS, J., HEGER, A., HOLM, L., SONNHAMMER, E. L. L., EDDY, S. R., BATEMAN, A. & FINN, R. D. 2012. The Pfam protein families database. *Nucleic Acids Research,* 40**,** D290-D301.

RADRICH, K., TSURUOKA, Y., DOBSON, P., GEVORGYAN, A., SWAINSTON, N., BAART, G. & SCHWARTZ, J.-M. 2010. Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Systems Biology,* 4**,** 114.

RAMAMOORTHY, V., ZHAO, X., SNYDER, A. K., XU, J.-R. & SHAH, D. M. 2007. Two mitogen-activated protein kinase signalling cascades mediate basal resistance to antifungal plant defensins in *Fusarium graminearum*. *Cellular Microbiology,* 9**,** 1491-1506.

REDDY, T. B. K., THOMAS, A. D., STAMATIS, D., BERTSCH, J., ISBANDI, M., JANSSON, J., MALLAJOSYULA, J., PAGANI, I., LOBOS, E. A. & KYRPIDES, N. C. 2014. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Research*.

REIS, H., PFIFFI, S. & HAHN, M. 2005. Molecular and functional characterization of a secreted lipase from *Botrytis cinerea*. *Molecular Plant Pathology,* 6**,** 257-267.

REMM, M., STORM, C. & SONNHAMMER, E. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol,* 314**,** 1041 - 1052.

RENTZSCH, R. & ORENGO, C. A. 2009. Protein function prediction – the power of multiplicity. *Trends in Biotechnology,* 27**,** 210-219.

RILEY, M. 1993. Functions of the gene products of *Escherichia coli*. *Microbiol Rev,* 57**,** 862-952.

RILEY, M. & LABEDAN, B. 1997. Protein evolution viewed through *Escherichia coli* Protein sequences: Introducing the notion of a structural segment of homology, the module. *Journal of Molecular Biology,* 268**,** 857-868.

ROGERS, L. M., FLAISHMAN, M. A. & KOLATTUKUDY, P. E. 1994. Cutinase gene disruption in *Fusarium solani* f sp *pisi* decreases its virulence on pea. *The Plant Cell,* 6**,** 935-45.

ROSSMANN, M. G., MORAS, D. & OLSEN, K. W. 1974. Chemical and biological evolution of a nucleotide-binding protein. *Nature,* 250**,** 194-199.

ROVENICH, H., BOSHOVEN, J. C. & THOMMA, B. P. H. J. 2014. Filamentous pathogen effector functions: of pathogens, hosts and microbiomes. *Current Opinion in Plant Biology,* 20**,** 96-103.

RUEPP, A., ZOLLNER, A., MAIER, D., ALBERMANN, K., HANI, J., MOKREJS, M., TETKO, I., GULDENER, U., MANNHAUPT, G., MUNSTERKOTTER, M. & MEWES, H. W. 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res,* 32**,** 5539-45.

SALWINSKI, L., MILLER, C. S., SMITH, A. J., PETTIT, F. K., BOWIE, J. U. & EISENBERG, D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res,* 32**,** D449-51.

SANGER, F. 1952. The Arrangement of Amino Acids in Proteins. *Advances in Protein Chemistry.* Academic Press.

SCHLEKER, S., GARCIA-GARCIA, J., KLEIN-SEETHARAMAN, J. & OLIVA, B. 2012. Prediction and Comparison of *Salmonella*-Human and *Salmonella-Arabidopsis* Interactomes. *Chemistry & Biodiversity,* 9**,** 991-1018.

SCHULTZ, J., MILPETZ, F., BORK, P. & PONTING, C. P. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A,* 95**,** 5857-64.

SEIDL, M. F., VAN DEN ACKERVEKEN, G., GOVERS, F. & SNEL, B. 2011. A Domain-Centric Analysis of Oomycete Plant Pathogen Genomes Reveals Unique Protein Organization. *Plant Physiology,* 155**,** 628-644.

SHANER, G., STROMBERG, E. L., LACY, G. H., BARKER, K. R. & PIRONE, T. P. 1992. Nomenclature and Concepts of Pathogenicity and Virulence. *Annual Review of Phytopathology,* 30**,** 47-66.

SILLITOE, I., CUFF, A. L., DESSAILLY, B. H., DAWSON, N. L., FURNHAM, N., LEE, D., LEES, J. G., LEWIS, T. E., STUDER, R. A., RENTZSCH, R., YEATS, C., THORNTON, J. M. & ORENGO, C. A. 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Research,* 41**,** D490-D498.

SILLITOE, I., LEWIS, T. E., CUFF, A., DAS, S., ASHFORD, P., DAWSON, N. L., FURNHAM, N., LASKOWSKI, R. A., LEE, D., LEES, J. G., LEHTINEN, S., STUDER, R. A., THORNTON, J. & ORENGO, C. A. 2015. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research,* 43**,** D376-D381.

SKAMNIOTI, P. & GURR, S. J. 2007. *Magnaporthe grisea* Cutinase2 Mediates Appressorium Differentiation and Host Penetration and Is Required for Full Virulence. *The Plant Cell,* 19**,** 2674-2689.

SKAMNIOTI, P. & GURR, S. J. 2008. Cutinase and hydrophobin interplay: A herald for pathogenesis? *Plant Signaling & Behavior,* 3**,** 248-250.

SMEDLEY, D., HAIDER, S., DURINCK, S., PANDINI, L., PROVERO, P., ALLEN, J., ARNAIZ, O., AWEDH, M. H., BALDOCK, R., BARBIERA, G., BARDOU, P., BECK, T., BLAKE, A., BONIERBALE, M., BROOKES, A. J., BUCCI, G., BUETTI, I., BURGE, S., CABAU, C., CARLSON, J. W., CHELALA, C., CHRYSOSTOMOU, C., CITTARO, D., COLLIN, O., CORDOVA, R., CUTTS, R. J., DASSI, E., GENOVA, A. D., DJARI, A., ESPOSITO, A., ESTRELLA, H., EYRAS, E., FERNANDEZ-BANET, J., FORBES, S., FREE, R. C., FUJISAWA, T., GADALETA, E., GARCIA-MANTEIGA, J. M., GOODSTEIN, D., GRAY, K., GUERRA-ASSUNÇÃO, J. A., HAGGARTY, B., HAN, D.-J., HAN, B. W., HARRIS, T., HARSHBARGER, J., HASTINGS, R. K., HAYES, R. D., HOEDE, C., HU, S., HU, Z.-L., HUTCHINS, L., KAN, Z., KAWAJI, H., KELIET, A., KERHORNOU, A., KIM, S., KINSELLA, R., KLOPP, C., KONG, L., LAWSON, D., LAZAREVIC, D., LEE, J.-H., LETELLIER, T., LI, C.-Y., LIO, P., LIU, C.-J., LUO, J., MAASS, A., MARIETTE, J., MAUREL, T., MERELLA, S., MOHAMED, A. M., MOREEWS, F., NABIHOUDINE, I., NDEGWA, N., NOIROT, C., PEREZ-LLAMAS, C., PRIMIG, M., QUATTRONE, A., QUESNEVILLE, H., RAMBALDI, D., REECY, J., RIBA, M., ROSANOFF, S., SADDIQ, A. A., SALAS, E., SALLOU, O., SHEPHERD, R., SIMON, R., SPERLING, L., SPOONER, W., STAINES, D. M., STEINBACH, D., STONE, K., STUPKA, E., TEAGUE, J. W., DAYEM ULLAH, A. Z., WANG, J., WARE, D., et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research,* 43**,** W589-W598.

SMITH, A. M., COUPLAND, G., DOLAN, L., HARBERD, N., JONES, J., MARTIN, C., SABLOWSKI, R. & AMEY, A. 2010. *Plant Biology*.

SOBROVA, P., ADAM, V., VASATKOVA, A., BEKLOVA, M., ZEMAN, L. & KIZEK, R. 2010. Deoxynivalenol and its toxicity. *Interdisciplinary Toxicology,* 3**,** 94-99.

SON, H., SEO, Y.-S., MIN, K., PARK, A. R., LEE, J., JIN, J.-M., LIN, Y., CAO, P., HONG, S.-Y., KIM, E.-K., LEE, S.-H., CHO, A., LEE, S., KIM, M.-G., KIM, Y., KIM, J.-E., KIM, J.-C., CHOI, G. J., YUN, S.-H., LIM, J. Y., KIM, M., LEE, Y.-H., CHOI, Y.-D. & LEE, Y.-W. 2011. A Phenome-Based Functional Analysis of Transcription Factors in the Cereal Head Blight Fungus, *Fusarium graminearum*. *PLoS Pathog,* 7**,** e1002310.

SORRELLS, TREVOR R. & JOHNSON, ALEXANDER D. 2015. Making Sense of Transcription Networks. *Cell,* 161**,** 714-723.

SPERSCHNEIDER, J., DODDS, P. N., GARDINER, D. M., MANNERS, J. M., SINGH, K. B. & TAYLOR, J. M. 2015. Advances and Challenges in Computational Prediction of Effectors from Plant Pathogenic Fungi. *PLOS Pathogens,* 11**,** e1004806.

SPERSCHNEIDER, J., GARDINER, D. M., TAYLOR, J. M., HANE, J. K., SINGH, K. B. & MANNERS, J. M. 2013. A comparative hidden Markov model analysis pipeline identifies proteins characteristic of cereal-infecting fungi. *BMC Genomics,* 14**,** 807-807.

STAHL, D. J. & SCHÄFER, W. 1992. Cutinase is not required for fungal pathogenicity on pea. *The Plant Cell,* 4**,** 621-9.

STRIEKER, M., TANOVIĆ, A. & MARAHIEL, M. A. 2010. Nonribosomal peptide synthetases: structures and dynamics. *Current Opinion in Structural Biology,* 20**,** 234-240.

STUART, J., SEGAL, E., KOLLER, D. & KIM, S. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science,* 302**,** 249 - 255.

SWEIGARD, J., CHUMLEY, F. & VALENT, B. 1992. Disruption of a *Manaporthe grisea* cutinase gene. *Molecular and General Genetics MGG,* 232**,** 183-190.

TAUTZ, D. & DOMAZET-LOŠO, T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet,* 12**,** 692-702.

TELESFORD, Q. K., JOYCE, K. E., HAYASAKA, S., BURDETTE, J. H. & LAURIENTI, P. J. 2011. The Ubiquity of Small-World Networks. *Brain Connectivity,* 1**,** 367-375.

THAKUR, K., CHAWLA, V., BHATTI, S., SWARNKAR, M. K., KAUR, J., SHANKAR, R. & JHA, G. 2013. De novo transcriptome sequencing and analysis for Venturia inaequalis, the devastating apple scab pathogen. *PLoS One,* 8**,** e53937.

TUCKER, S. L., BESI, M. I., GALHANO, R., FRANCESCHETTI, M., GOETZ, S., LENHERT, S., OSBOURN, A. & SESMA, A. 2010. Common genetic pathways regulate organ-specific infection-related development in the rice blast fungus. *Plant Cell,* 22**,** 953-72.

URBAN, M., IRVINE, A. G., CUZICK, A. & HAMMOND-KOSACK, K. E. 2015a. Using the Pathogen-Host Interactions database (PHI-base) to investigate plant pathogen genomes and genes implicated in virulence. *Frontiers in Plant Science,* 6.

URBAN, M., MOTT, E., FARLEY, T. & HAMMOND-KOSACK, K. 2003. The *Fusarium graminearum* MAP1 gene is essential for pathogenicity and development of perithecia. *Molecular Plant Pathology,* 4**,** 347-359.

URBAN, M., PANT, R., RAGHUNATH, A., IRVINE, A. G., PEDRO, H. & HAMMOND-KOSACK, K. E. 2015b. The Pathogen-Host Interactions database (PHI-base): additions and future developments. *Nucleic Acids Research,* 43**,** D645-D655.

VAN DONGEN, S. 2000. *Graph Clustering by Flow Simulation.* PhD, University of Utrecht.

VARGAS, W. A., MARTÍN, J. M. S., RECH, G. E., RIVERA, L. P., BENITO, E. P., DÍAZ-MÍNGUEZ, J. M., THON, M. R. & SUKNO, S. A. 2012. Plant Defense Mechanisms Are Activated during Biotrophic and Necrotrophic Development of *Colletotricum graminicola* in Maize. *Plant Physiology,* 158**,** 1342-1358.

VINCENT, D. B., JEAN-LOUP, G., RENAUD, L. & ETIENNE, L. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment,* 2008**,** P10008.

VOGEL, C., BASHTON, M., KERRISON, N., CHOTHIA, C. & TEICHMANN, S. 2004a. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology,* 14**,** 208 - 216.

VOGEL, C., BERZUINI, C., BASHTON, M., GOUGH, J. & TEICHMANN, S. A. 2004b. Supra-domains: evolutionary units larger than single protein domains. *J Mol Biol,* 336**,** 809-23.

WALHOUT, A. J. M. & VIDAL, M. 2001. Protein interaction maps for model organisms. *Nat Rev Mol Cell Biol,* 2**,** 55-63.

WANG, C., ZHANG, S., HOU, R., ZHAO, Z., ZHENG, Q., XU, Q., ZHENG, D., WANG, G., LIU, H., GAO, X., MA, J. W., KISTLER, H. C., KANG, Z. & XU, J. R. 2011a. Functional analysis of the kinome of the wheat scab fungus *Fusarium graminearum*. *PLoS Pathog,* 7**,** e1002460.

WANG, T. Y., HE, F., HU, Q. W. & ZHANG, Z. 2011b. A predicted protein-protein interaction network of the filamentous fungus *Neurospora crassa*. *Mol Biosyst,* 7**,** 2278-85.

WANG, Y.-C., LIN, C., CHUANG, M.-T., HSIEH, W.-P., LAN, C.-Y., CHUANG, Y.-J. & CHEN, B.-S. 2013. Interspecies protein-protein interaction network construction for characterization of host-pathogen interactions: a *Candida albicans*-zebrafish interaction study. *BMC Systems Biology,* 7**,** 79.

WANG, Z.-Y., JENKINSON, J. M., HOLCOMBE, L. J., SOANES, D. M., VENEAULT-FOURREY, C., BHAMBRA, G. K. & TALBOT, N. J. 2005. The molecular biology of appressorium turgor generation by the rice blast fungus *Magnaporthe grisea*.

WANG, Z., ZHANG, X., LE, M., XU, D., STACEY, G. & CHENG, J. 2011c. A protein domain co-occurrence network approach for predicting protein function and inferring species phylogeny. *PLoS One,* 6**,** e17906.

WARREN, R. L., SUTTON, G. G., JONES, S. J. M. & HOLT, R. A. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics,* 23**,** 500-501.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.

WEISSMAN, K. J. & LEADLAY, P. F. 2005. Combinatorial biosynthesis of reduced polyketides. *Nat Rev Micro,* 3**,** 925-936.

WETLAUFER, D. B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A,* 70**,** 697-701.

WILD, C. P. & GONG, Y. Y. 2010. Mycotoxins and human disease: a largely ignored global health issue. *Carcinogenesis,* 31**,** 71-82.

WILSON, D., MADERA, M., VOGEL, C., CHOTHIA, C. & GOUGH, J. 2007a. The SUPERFAMILY database in 2007: families and functions. *Nucl Acids Res,* 35**,** D308 - D313.

WILSON, D., PETHICA, R., ZHOU, Y., TALBOT, C., VOGEL, C., MADERA, M., CHOTHIA, C. & GOUGH, J. 2009. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research,* 37**,** D380-D386.

WILSON, D., TUTULAN-CUNITA, A., JUNG, W., HAUSER, N. C., HERNANDEZ, R., WILLIAMSON, T., PIEKARSKA, K., RUPP, S., YOUNG, T. & STATEVA, L. 2007b. Deletion of the high-affinity cAMP phosphodiesterase encoded by PDE2 affects stress responses and virulence in *Candida albicans*. *Mol Microbiol,* 65**,** 841-56.

WILSON, R. A. & TALBOT, N. J. 2009. Under pressure: investigating the biology of plant infection by *Magnaporthe oryzae*. *Nat Rev Micro,* 7**,** 185-195.

WINNENBURG, R., BALDWIN, T., URBAN, M., RAWLINGS, C., KOHLER, J. & HAMMOND-KOSACK, K. 2006. PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res***,** D459 - 464.

WOGAN, G. N., HECHT, S. S., FELTON, J. S., CONNEY, A. H. & LOEB, L. A. 2004. Environmental and chemical carcinogenesis. *Semin Cancer Biol,* 14**,** 473-86.

WOLKOW, P. M., SISLER, H. D. & VIGIL, E. L. 1983. Effect of inhibitors of melanin biosynthesis on structure and function of appressoria of *Colletotrichum lindemuthianum*. *Physiological Plant Pathology,* 23**,** 55-71.

WONG, P., WALTER, M., LEE, W., MANNHAUPT, G., MUNSTERKOTTER, M., MEWES, H. W., ADAM, G. & GULDENER, U. 2011. FGDB: revisiting the genome annotation of the plant pathogen *Fusarium graminearum*. *Nucleic Acids Res,* 39**,** D637-9.

WU, C. H., YEH, L. S., HUANG, H., ARMINSKI, L., CASTRO-ALVEAR, J., CHEN, Y., HU, Z., KOURTESIS, P., LEDLEY, R. S., SUZEK, B. E., VINAYAKA, C. R., ZHANG, J. & BARKER, W. C. 2003. The Protein Information Resource. *Nucleic Acids Res,* 31**,** 345-7.

WU, X., ZHU, L., GUO, J., ZHANG, D. & LIN, K. 2006. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res,* 34**,** 2137 - 2150.

WUCHTY, S. 2001. Scale-Free Behavior in Protein Domain Networks. *Molecular Biology and Evolution,* 18**,** 1694-1702.

XIA, J., WANG, S. & LEI, Y. 2010. Computational methods for the prediction of protein-protein interactions. *Protein Pept Lett,* 17**,** 1069 - 1078.

XIE, X., JIN, J. & MAO, Y. 2011. Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks. *BMC Evolutionary Biology,* 11**,** 242.

XU, J., PENG, Y., DICKMAN, M. & SHARON, A. 2006. The dawn of fungal pathogen genomics. *Annu Rev Phytopathol,* 44**,** 337 - 366.

XUE, C., PARK, G., CHOI, W., ZHENG, L., DEAN, R. A. & XU, J.-R. 2002. Two novel fungal virulence genes specifically expressed in appressoria of the rice blast fungus. *The Plant cell,* 14**,** 2107-2119.

YEATS, C. A. & ORENGO, C. A. 2001. Evolution of Protein Domains. *eLS.* John Wiley & Sons, Ltd.

YELLABOINA, S., TASNEEM, A., ZAYKIN, D., RAGHAVACHARI, B. & JOTHI, R. 2011. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res,* 39**,** D730 - D735.

YU, H., LUSCOMBE, N. M., LU, H. X., ZHU, X., XIA, Y., HAN, J. D., BERTIN, N., CHUNG, S., VIDAL, M. & GERSTEIN, M. 2004. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res,* 14**,** 1107-18.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research,* 18**,** 821-829.

ZHANG, H., TANG, W., LIU, K., HUANG, Q., ZHANG, X., YAN, X., CHEN, Y., WANG, J., QI, Z., WANG, Z., ZHENG, X., WANG, P. & ZHANG, Z. 2011. Eight RGS and RGS-like proteins orchestrate growth, differentiation, and pathogenicity of *Magnaporthe oryzae*. *PLoS Pathog,* 7**,** e1002450.

ZHANG, Y., ZHANG, K., FANG, A., HAN, Y., YANG, J., XUE, M., BAO, J., HU, D., ZHOU, B., SUN, X., LI, S., WEN, M., YAO, N., MA, L. J., LIU, Y., ZHANG, M., HUANG, F., LUO, C., ZHOU, L., LI, J., CHEN, Z., MIAO, J., WANG, S., LAI, J., XU, J. R., HSIANG, T., PENG, Y. L. & SUN, W. 2014. Specific adaptation of *Ustilaginoidea virens* in occupying host florets revealed by comparative and functional genomics. *Nat Commun,* 5**,** 3849.

ZHAO, B., SI, H. L., SUN, Z. Y., XU, Z., CHEN, Z., ZHANG, J. L., XING, J. H. & DONG, J. G. 2015. Identification of Development and Pathogenicity Related Gene in *Botrytis cinerea* via Digital Gene Expression Profile. *Jundishapur Journal of Microbiology,* 8**,** e22432.

ZHAO, X. M., ZHANG, X. W., TANG, W. H. & CHEN, L. 2009. FPPI: *Fusarium graminearum* protein-protein interaction database. *J Proteome Res,* 8**,** 4714-21.

ZHENG, D., ZHANG, S., ZHOU, X., WANG, C., XIANG, P., ZHENG, Q. & XU, J.-R. 2012. The *FgHOG1* Pathway Regulates Hyphal Growth, Stress Responses, and Plant Infection in *Fusarium graminearum*. *PLoS ONE,* 7**,** e49495.

ZHOU, H., REZAEI, J., HUGO, W., GAO, S., JIN, J., FAN, M., YONG, C.-H., WOZNIAK, M. & WONG, L. 2013. Stringent DDI-based Prediction of *H. sapiens-M. tuberculosis* H37Rv Protein-Protein Interactions. *BMC Systems Biology,* 7**,** 1-15.

ZHU, P., GU, H., JIAO, Y., HUANG, D. & CHEN, M. 2011. Computational identification of protein-protein interactions in rice based on the predicted rice interactome network. *Genomics Proteomics Bioinformatics,* 9**,** 128-37.

ZIELONKA, L., WASKIEWICZ, A., BESZTERDA, M., KOSTECKI, M., DABROWSKI, M., OBREMSKI, K., GOLINSKI, P. & GAJECKI, M. 2015. Zearalenone in the Intestinal Tissues of Immature Gilts Exposed per os to Mycotoxins. *Toxins (Basel),* 7**,** 3210-23.

ZINEDINE, A., SORIANO, J. M., MOLTÓ, J. C. & MAÑES, J. 2007. Review on the toxicity, occurrence, metabolism, detoxification, regulations and intake of zearalenone: An oestrogenic mycotoxin. *Food and Chemical Toxicology,* 45**,** 1-18.

ZIPFEL, C. 2009. Early molecular events in PAMP-triggered immunity. *Current Opinion in Plant Biology,* 12**,** 414-420.

Appendix A.

Additional tables to Chapter 3

Table A-1 Plant pathogen species in PHI-base version 3.2.
Here species with experimentally verified pathogenicity, virulence, or effector genes in PHI-base version 3.2 are listed. The third column shows the number of hosts with which a pathogen (column one) interacts. The fourth column lists the number of experimentally verified genes per pathogen.

Pathogen Name	Taxonomy	No of hosts	No of genes
Magnaporthe grisea	Ascomycota	4	161
Ustilago maydis	Basidiomycota	1	137
Fusarium graminearum	Ascomycota	7	65
Pseudomonas syringae	Bacteria	7	50
Botrytis cinerea	Ascomycota	14	42
Fusarium oxysporum	Ascomycota	5	27
Mycosphaerella graminicola	Ascomycota	1	22
Claviceps purpurea	Ascomycota	1	16
Phytophthora infestans	Oomycetes	5	15
Stagonospora nodorum	Ascomycota	2	15
Cochliobolus carbonum	Ascomycota	1	14
Colletotrichum lagenarium	Ascomycota	1	11
Leptosphaeria maculans	Ascomycota	3	9
Colletotrichum gloeosporioides	Ascomycota	4	8
Cladosporium fulvum	Ascomycota	1	7
Cochliobolus heterostrophus	Ascomycota	1	7
Colletotrichum lindemuthianum	Ascomycota	1	7
Cryphonectria parasitica	Ascomycota	2	7
Nectria haematococca (related*: Fusarium solani*)	Ascomycota	2	7
Phytophthora sojae	Oomycetes	2	6
Alternaria alternata	Ascomycota	3	5
Burkholderia glumae	Bacteria	1	5
Melampsora lini	Basidiomycota	2	5
Alternaria brassicicola	Ascomycota	6	4
Colletotrichum graminicola	Ascomycota	1	4
Hyaloperonospora parasitica	Oomycetes	3	4
Blumeria graminis	Ascomycota	2	3
Epichloe festucae	Ascomycota	1	3
Streptomyces turgidiscabies	Bacteria	3	3
Botrytis elliptica	Ascomycota	1	2
Cercospora nicotianae	Ascomycota	1	2
Cercospora zeae-maydis	Ascomycota	1	2
Colletotrichum trifolii	Ascomycota	1	2
Gibberella moniliformis	Ascomycota	2	2
Rhynchosporium secalis	Ascomycota	1	2
Ustilago hordei	Basidiomycota	1	2

Table A-2 Animal pathogen species in PHI-base version 3.2.
Here species with experimentally verified pathogenicity, virulence, or effector genes in PHI-base version 3.2 are listed. The third column shows the number of hosts with which a pathogen (column one) interacts. The fourth column lists the number of experimentally verified genes per pathogen. *Model organism.

Pathogen name	Taxonomy	No of hosts	No of genes
Candida albicans	Ascomycota	5	126
Cryptococcus neoformans	Basidiomycota	8	82
Salmonella enterica	Bacteria	2	63
Aspergillus fumigatus	Ascomycota	2	23
Vibrio cholerae	Bacteria	3	14
*Saccharomyces cerevisiae**	Ascomycota	1	7
Wangiella (Exophiala) dermatitidis	Ascomycota	1	4
Candida glabrata	Ascomycota	2	4
Cryptococcus gattii	Basidiomycota	1	3
Histoplasma capsulatum	Ascomycota	1	3
Flavobacterium psychrophilum	Bacteria	2	2
Trichophyton rubrum	Ascomycota	1	2
Aspergillus nidulans	Ascomycota	1	1
Beauveria bassiana	Ascomycota	1	1
Blastomyces dermatitidis	Ascomycota	1	1
Candida tropicalis	Ascomycota	1	1
Coccidioides immitis	Ascomycota	1	1
Coccidioides posadasii	Ascomycota	1	1

Appendix B.

Additional figures and tables to Chapter 4

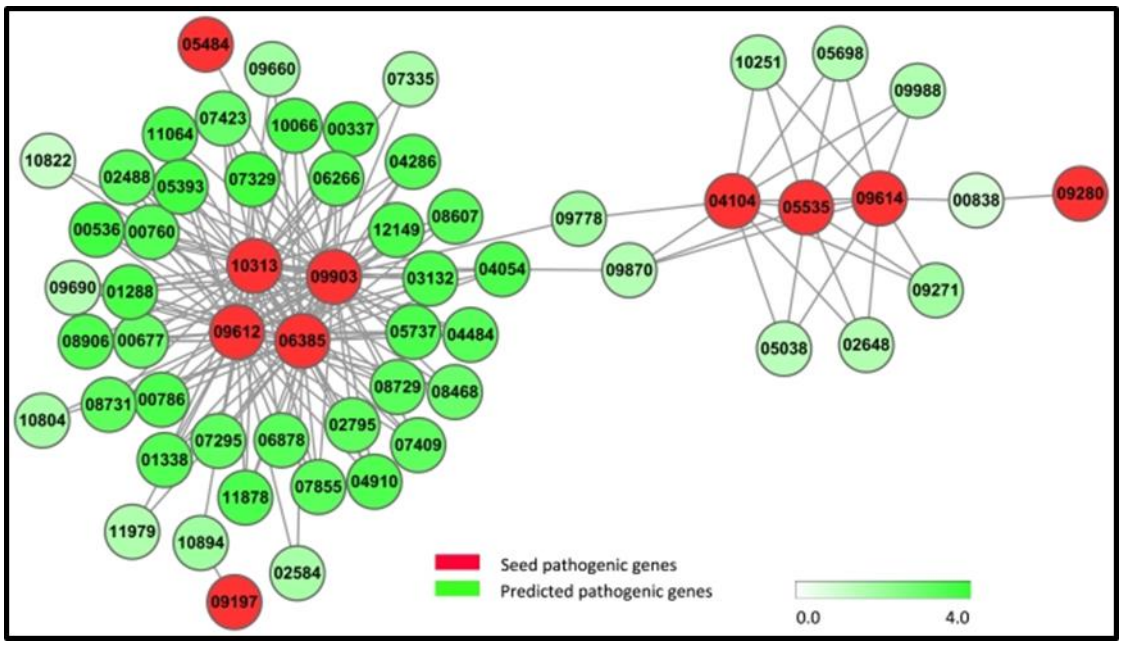


Figure B-1 The filtered pathogenic network (Liu et al., 2010).
The red vertices denote 'seed' genes, the green vertices are genes that interact with at least two 'seed' genes, and each vertex is assigned a weight. The colour bar represents the relationship between colour and weight, where the deeper the colour is the larger the weight is. For clarity, the prefix "FGSG_" was omitted from gene names.

Table B-1 *F. graminearum* genes present in PHI-base version 3.1 (release date: 1st Apr 2008) - detailed table.
This data was used in study by Liu et al. (2010). Please note the authors failed to dissect the genes according to the phenotypes affecting pathogenicity and those not affecting pathogenicity and used all genes to predict the *F. graminearum* candidate genes for the pathogenicity.

No	PHI-base ID	FGSG ID	Gene name	Phenotype of mutant
1	PHI:44	FGSG_03537	TRI5	reduced virulence/unaffected pathogenicity
2	PHI:266	FGSG_10313	MGV1	reduced virulence/ loss of pathogenicity
3	PHI:304	FGSG_06631	GzCPS1	reduced virulence
4	PHI:309	FGSG_06385	MAP1 (related: GPMK1)	reduced virulence/ loss of pathogenicity
5	PHI:355	FGSG_05658	GzmetE	reduced virulence
6	PHI:432	FGSG_05906	FGL1	reduced virulence
7	PHI:439	FGSG_03536	TRI6	reduced virulence
8	PHI:442	FGSG_10825	MSY1	reduced virulence
9	PHI:443	FGSG_01932	CBL1	reduced virulence
10	PHI:444	FGSG_01555	ZIF1	reduced virulence
11	PHI:445	FGSG_00376	NOS1	reduced virulence
12	PHI:446	FGSG_00332	TBL1	reduced virulence
13	PHI:525	FGSG_03543	TRI14	reduced virulence
14	PHI:712		GIP1	unaffected pathogenicity
15	PHI:713	FGSG_02395	PKS13 (ZEA2)	unaffected pathogenicity
16	PHI:714	FGSG_12126	PKS4 (ZEA1)	unaffected pathogenicity
17	PHI:715	FGSG_02398	ZEB2	unaffected pathogenicity
18	PHI:716		ZEB1	unaffected pathogenicity
19	PHI:717	FGSG_04488	PLSP1	unaffected pathogenicity
20	PHI:718		PKS6	unaffected pathogenicity
21	PHI:719		GzFUS1	unaffected pathogenicity
22	PHI:720	FGSG_02324	AUR1	unaffected pathogenicity
23	PHI:721		GRS1	unaffected pathogenicity
24	PHI:722		PGL1	unaffected pathogenicity
25	PHI:723		PKS1	unaffected pathogenicity
26	PHI:724		PKS9	unaffected pathogenicity
27	PHI:725		PKS11	unaffected pathogenicity
28	PHI:726		PKS7	unaffected pathogenicity
29	PHI:727		PKS17	unaffected pathogenicity
30	PHI:728		PKS5	unaffected pathogenicity
31	PHI:729		PKS2	unaffected pathogenicity
32	PHI:730		KSA1	unaffected pathogenicity
33	PHI:731	FGSG_01665	FgFSR1	reduced virulence
34	PHI:733	FGSG_02095	FBP1	reduced virulence
35	PHI:743	FGSG_01939	ARG2	reduced virulence
36	PHI:744	FGSG_02506	ADE5	reduced virulence
37	PHI:861	FGSG_10114	RAS2	reduced virulence
38	PHI:1002	FGSG_05955	GCS1	reduced virulence
39	PHI:1004	FGSG_09903	STE7	loss of pathogenicity
40	PHI:1005	FGSG_09612	FgHOG1	reduced virulence/ unaffected pathogenicity
41	PHI:1006	FGSG_09197	HMR1	reduced virulence
42	PHI:1007	FGSG_03747	NPS6	reduced virulence

Table B- 1 continued

No	PHI-base ID	FGSG ID	Gene name	Phenotype of mutant
43	PHI:1010	FGSG_05371	SID1	reduced virulence
44	PHI:1011		FET3	unaffected pathogenicity
45	PHI:1012		GzGPA1	unaffected pathogenicity
46	PHI:1013		GzGPA2	reduced virulence
47	PHI:1014		GzGPA3	unaffected pathogenicity
48	PHI:1015		GzGPB1	reduced virulence
49	PHI:1016	FGSG_05484	STE11	loss of pathogenicity

Table B-2 List of *F. graminearum* genes in PHI-base version 3.2 (release date: 14th Dec 2009) that were experimentally proven to cause loss of pathogenic activity when disrupted or deleted.
This data was used in Chapter 4, section 4.3 (Prediction and characterisation *of F. graminearum* candidate genes).

No	FGSG_No	gene name	Function	PHI-base ID
1	FGSG_05484	*STE11*	MAPKKK; hypersensitive to MsDEF1	PHI:1016
2	FGSG_06385	*MAP1(gpmk1)*	MAPK pathogenicity	PHI:309
3	FGSG_09903	*STE7*	MAPKK; hypersensitive to MsDEF1	PHI:1004

Table B-3 List of *F. graminearum* genes in PHI-base version 3.2 (release date: 14th Dec 2009) that were experimentally proven to increase virulence when disrupted or deleted, i.e. they act as repressors of virulence in wild-type isolates
This data was used in Chapter 4, section 4.3 (Prediction and characterisation *of F. graminearum* candidate genes).

No	FGSG_No	gene name	Function	PHI-base ID
1	FGSG_00007	FGSG_00007	cytochrome P450 monooxygenase (DON biosynthesis); repressor for virulence	PHI:2393
2	FGSG_10397	FGSG_10397	unknown function; repressor for virulence	PHI:2394
3	FGSG_11025	*Tri15*	Putative transcription factor	PHI:1363

Table B-4 List of *F. graminearum* genes in PHI-base version 3.2 (release date: 14th Dec 2009) that were experimentally proven to reduce virulence or have no effect on pathogenicity (depending on the host) when disrupted or deleted
These results assign a "mixed outcome" phenotype. This data was used in Chapter 4, section 4.3 (Prediction and characterisation *of F. graminearum* candidate genes).

No	FGSG_No	gene name	function	PHI-base ID
1	FGSG_03537	*TRI5*	Trichodiene synthase	PHI:44
2	FGSG_05955	*GCS1*	Glycosylceramide synthase (Sphingolipid biosynthesis)	PHI:1002
3	FGSG_09612	*HOG1/Os-2*	MAPK osmo-sensing	PHI:1005

Table B-5 List of *F. graminearum* genes in PHI-base version 3.2 (release date: 14th Dec 2009) that were experimentally proven to reduce virulence when disrupted or deleted.
This data was used in Chapter 4, section 4.3 (Prediction and characterisation *of F. graminearum* candidate genes).

No	FGSG_No	gene name	function	PHI-base ID
1	FGSG_00332	*FTL1*	Transducin beta-subunit	PHI:446
2	FGSG_00376	*NOS1*	NADH:Ubiquinone oxidoreductase	PHI:445
3	FGSG_00950	*SYN1*	SNARE protein (transport docking and vesicle fusion)	
4	FGSG_01555	*ZIF1*	b-ZIP transcription factor	PHI:444
5	FGSG_01665	*FSR1*	Putative signalling scaffold protein	PHI:731
6	FGSG_01932	*CBL1*	Cystathionine beta-lyase	PHI:443
7	FGSG_01939	*ARG2*	Acetylglutamate synthase	PHI:743
8	FGSG_01964	*CHS5*	myosin-motor like chitinase	
9	FGSG_02095	*FBP1*	F-box protein involved in ubiquitin-mediated degradation	PHI:733
10	FGSG_02506	*ADE5*	Phosphoribosylamine-glycine ligase	PHI:744
11	FGSG_03536	*TRI6*	Transcription factor	PHI:439
12	FGSG_03538	*TRI10*	regulatory protein	PHI:2328
13	FGSG_03543	*TRI14*	putative trichodiene biosynthesis gene	PHI:525
14	FGSG_03747	*NPS6*	Non-ribosomal peptide synthetase for biosynthesis of extracellular siderophores	PHI:1007
15	FGSG_04104	*GPB1*	Guanine nucleotide-binding protein beta subunit	
16	FGSG_04111	*PTC1*	type 2C protein phosphatase	PHI:2326
17	FGSG_04355	*CID1*	Cyclin-C-like gene requried for infection and DON production1	PHI:2418, PHI:2419
18	FGSG_05371	*SID1*	Siderophore biosynthetic gene	PHI:1010
19	FGSG_05593	*MT2*	Sphingolipid C-9- methyltransferase	
20	FGSG_05658	*GzmetE*	Homoserine O-acetyltransferase	PHI:355
21	FGSG_05906	*FGL1*	Secreted Lipase	PHI:432
22	FGSG_06631	*CPS1*	adenylate-forming enzyme	PHI:304
23	FGSG_06680	*MES1*	role in cell-surface organisation	PHI:1078
24	FGSG_06874	*TOP1*	topoisomerase 1	PHI:1291
25	FGSG_09197	*HMR1*	3-hydroxy-3-methylglutaryl-coenzyme A reductase involved in isoprenoid biosynthesis	PHI:1006
26	FGSG_09614	*GPA2*	Guanine nucleotide-binding protein alpha-3 subunit	
27	FGSG_09895	*NTH1*	neutral trehalase	
28	FGSG_09897	*SNF1*	sucrose nonfermenting protein kinase	PHI:1197
29	FGSG_09907	*FCV1*	conserved hypothetical protein	
30	FGSG_09908	*PKAR*	Protein kinase A regulatory subunit	
31	FGSG_09928	*SYN2*	SNARE protein (transport docking and vesicle fusion)	
32	FGSG_10114	*RAS2*	Ras GTPase	PHI:861
33	FGSG_10313	*MGV1*	MAP kinase; essential for female fertility and heterokaryon formation	PHI:266
34	FGSG_10825	*MSY1*	Methionine synthase	PHI:442
35	FGSG_12039	*CHS7*	myosin-motor like chitinase	

Table B-6 List of *F. graminearum* genes in PHI-base version 3.2 (release date: 14th Dec 2009) that were experimentally proven to have no effect on pathogenicity when disrupted or deleted.
This data was used in Chapter 4, section 4.3 (Prediction and characterisation *of F. graminearum* candidate genes).

No	FGSG_No	gene name	function	PHI-base ID
1	FGSG_01364	*CCH1*	related to voltage gated Ca2+ channel	PHI:1080
2	FGSG_01790	*PKS11*	polyketide synthase	PHI:725
3	FGSG_02324	*AUR1*	polyketide synthase that catalyse the condensation of one acetyl-CoA and six malonyl-CoA resulting in formation of nor-rubrofusarin	PHI:720
4	FGSG_02328	*GIP1*	laccase that catalyse the dimerization of two 9-hydroxyrubrofusarin in C7 positions	PHI:712
5	FGSG_02395	*PKS13*	polyketide synthase	PHI:713
6	FGSG_02398	*ZEB2*	conserved hypothetical protein	PHI:715
7	FGSG_03340	*PKS17*	polyketide synthase	PHI:727
8	FGSG_03964	*GRS1*	polyketide synthase	PHI:721
9	FGSG_04488	*PLSP1*	conserved hypothetical protein	PHI:717
10	FGSG_04510	*PHI:1087*	related to monophenol monooxygenase (tyrosinase)	PHI:1087
11	FGSG_04610	*PHI:1094*	related to alpha-glucoside transport protein	PHI:1094
12	FGSG_04694	*PKS2*	polyketide synthase	PHI:729
13	FGSG_05371	*SID1*	related to L-ornithine N5-hydroxylase	
14	FGSG_05535	*GPA1*	probable G protein alpha chain	PHI:76
15	FGSG_05794	*PKS5*	polyketide synthase	PHI:728
16	FGSG_07062	*PHI:1096*	related to ERD1 protein, required for retention of luminal ER proteins	PHI:1096
17	FGSG_07226	*KSA1*	probable CEM1 - beta-keto-acyl-ACP synthase, mitochondrial	PHI:730
18	FGSG_07798	*FUS1*	probable polyketide synthase	
19	FGSG_08208	*PKS6*	polyketide synthase	PHI:718
20	FGSG_08695	*PLS1*	conserved hypothetical protein	PHI:1079
21	FGSG_08737	*PHI:1091*	probable woronin body major protein precursor	PHI:1091
22	FGSG_08795	*PKS7*	polyketide synthase	PHI:726
23	FGSG_09182	*PGL1*	hypothetical protein similar to polyketide synthase	PHI:722
24	FGSG_09759	*PHI:1088*	related to zinc/cadmium resistance protein	PHI:1088
25	FGSG_09891	*AB1*	probable arsenite translocating ATPase (ASNA1)	
26	FGSG_09893	*AB2*	probable cytosolic nonspecific dipeptidase	
27	FGSG_09896	*ICL1*	probable isocitrate lyase (acu-3)	
28	FGSG_09900	*AB3*	conserved hypothetical protein	
29	FGSG_09905	*AB4*	hypothetical protein	
30	FGSG_09906	*AB5*	conserved hypothetical protein	
31	FGSG_09988	*GPA3*	probable G protein alpha chain	
32	FGSG_10464	*PKS9*	polyketide synthase	PHI:724
33	FGSG_10548	*PKS1*	polyketide synthase	
34	FGSG_12126	*PKS4*	polyketide synthase	PHI:714

Table B-7 Main functional categories within the FunCat scheme (source: MIPS).
This data was used in Chapter 4, sections 4.3.1 (Functional characterisation of *Fusarium graminearum* genes using FunCat ontology) and 4.3.2 (Using network for prediction, Table 4-9 and Figure 4-6).

No	MIPS Functional Category
01	METABOLISM
02	ENERGY
04	STORAGE PROTEIN
10	CELL CYCLE AND DNA PROCESSING
11	TRANSCRIPTION
12	PROTEIN SYNTHESIS
14	PROTEIN FATE
16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
18	PROTEIN ACTIVITY REGULATION
20	CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES
30	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM
32	CELL RESCUE, DEFENSE AND VIRULENCE
34	INTERACTION WITH THE CELLULAR ENVIRONMENT
36	INTERACTION WITH THE ENVIRONMENT (Systemic)
38	TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS
40	CELL FATE
41	DEVELOPMENT (Systemic)
42	BIOGENESIS OF CELLULAR COMPONENTS
43	CELL TYPE DIFFERENTIATION
45	TISSUE DIFFERENTIATION
47	ORGAN DIFFERENTIATION
70	SUBCELLULAR LOCALIZATION
73	CELL TYPE LOCALIZATION
75	TISSUE LOCALIZATION
77	ORGAN LOCALIZATION
78	UBIQUITOUS EXPRESSION
98	CLASSIFICATION NOT YET CLEAR-CUT
99	UNCLASSIFIED PROTEINS

Table B-8 List of *M. oryzae* genes that were experimentally proven to loss of pathogenic activity when disrupted or deleted.
Data from PHI-base version 3.2 (release date: 14th Dec 2009). This data was used in Chapter 4, section 4.4 (Prediction of candidate genes in *Magnaporthe oryzae* – a pilot study).

No	MGG ID	gene name	function	PHI-base ID
1	MGG_06368	CPKA	cAMP-dependent protein kinase	PHI:36
2	MGG_00883		mitogen activated protein kinase3	PHI:777
3	MGG_00435		integral membrane protein	PHI:874
4	MGG_04587		mediator of RNA polymerase II transcription subunit 16	PHI:875
5	MGG_10323	MgRho3	GTP-binding protein rho3	PHI:1061
6	MGG_01204	MIG1	MADS-box MEF2 type transcription factor	PHI:1070

Table B-9 List of *M. oryzae* genes that were experimentally proven to loss of pathogenic activity or reduced virulence when disrupted or deleted.
These results assign a "mixed outcome" phenotype. Data from PHI-base version 3.2 (release date: 14th Dec 2009). This data was used in Chapter 4, section 4.4 (Prediction of candidate genes in *Magnaporthe oryzae* – a pilot study).

No	MGG ID	gene name	function	PHI-base ID
1	MGG_01481		peroxisomal targeting signal 2 receptor	PHI:772 / PHI:797
2	MGG_09250		gamma-butyrobetaine dioxygenase	PHI:774 / PHI:792
3	MGG_02423		ER lumen protein retaining receptor 2	PHI:782 /PHI:796
4	MGG_02731		cell division control protein 42	PHI:780 / PHI:787, PHI:808

Table B-10 List of *M. oryzae* genes that were experimentally proven to reduce virulence activity when disrupted or deleted.
Data from PHI-base version 3.2 (release date: 14th Dec 2009). This data was used in Chapter 4, section 4.4 (Prediction of candidate genes in *Magnaporthe oryzae* – a pilot study).

No	MGG ID	gene name	function	PHI-base ID
1	MGG_09898	MAC1	adenylate cyclase	PHI:81
2	MGG_00365	MAGB	guanine nucleotide-binding protein subunit alpha	PHI:83
3	MGG_00527	EMP1	hypothetical protein	PHI:350
4	MGG_01173	MHP1	hydrophobin	PHI:458
5	MGG_12175	SSM1	tyrocidine synthetase 1	PHI:739
6	MGG_11899		SH3 domain-containing protein	PHI:773
7	MGG_09471		neutral trehalase	PHI:775 / PHI:794
8	MGG_00692		cell pattern formation-associated protein stuA	PHI:776 / PHI:802
9	MGG_14719		conserved hypothetical protein	PHI:781
10	MGG_04163		conserved hypothetical protein	PHI:783
11	MGG_00056		xanthoxin dehydrogenase	PHI:784
12	MGG_04128		DUF1237 domain-containing protein	PHI:785
13	MGG_04538		conserved hypothetical protein	PHI:786
14	MGG_04116		SH3 domain-containing protein	PHI:789
15	MGG_07259		conserved hypothetical protein	PHI:790
16	MGG_13024		hypothetical protein	PHI:793
17	MGG_03451		conserved hypothetical protein	PHI:795
18	MGG_12026		predicted protein	PHI:798
19	MGG_03530		conserved hypothetical protein	PHI:799
20	MGG_13324		conserved hypothetical protein	PHI:800
21	MGG_04621		conserved hypothetical protein	PHI:801
22	MGG_04629		integral membrane protein	PHI:803
23	MGG_08628		3'-5'exoribonuclease CSL4	PHI:804
24	MGG_00124		conserved hypothetical protein	PHI:805
25	MGG_04137		CTLH domain-containing protein	PHI:806
26	MGG_06951		CAAX prenyl protease 1	PHI:807
27	MGG_02436		conserved hypothetical protein	PHI:810
28	MGG_10510		ribonuclease T2	PHI:811
29	MGG_10702		conserved hypothetical protein	PHI:812
30	MGG_04685		conserved hypothetical protein	PHI:815
31	MGG_04582		conserved hypothetical protein	PHI:816
32	MGG_02049		interferon-induced GTP-binding protein Mx2	PHI:817
33	MGG_07061		conserved hypothetical protein	PHI:819
34	MGG_07075	MSP1	ATPase family AAA domain-containing protein 1-A	PHI:860
35	MGG_03284		DNA mismatch repair protein	PHI:872
36	MGG_02443		conserved hypothetical protein	PHI:873
37	MGG_15116		conserved hypothetical protein	PHI:876

Table B-10 continues

No	MGG ID	gene name	function	PHI-base ID
38	MGG_00383		S-adenosylmethionine synthetase	PHI:877
39	MGG_12142		predicted protein	PHI:878
40	MGG_04556		alcohol dehydrogenase 1	PHI:881
41	MGG_04985		conserved hypothetical protein	PHI:882
42	MGG_08560		predicted protein	PHI:883
43	MGG_02240		conserved hypothetical protein	PHI:885
44	MGG_05174		pinin/SDK/memA domain-containing protein	PHI:887
45	MGG_01707		conserved hypothetical protein	PHI:888
45	MGG_09263		C6 zinc finger domain-containing protein	PHI:889
47	MGG_07015		DNA repair protein Rad7	PHI:890
48	MGG_01748		conserved hypothetical protein	PHI:891
49	MGG_02986		DNA polymerase zeta catalytic subunit	PHI:893
50	MGG_00803	MoSNF1	carbon catabolite-derepressing protein kinase	PHI:1058

Table B-11 List of *M. oryzae* genes that were experimentally proven not to have effect on pathogenicity when disrupted or deleted.
Data from PHI-base version 3.2 (release date: 14th Dec 2009). This data was used in Chapter 4, section 4.4 (Prediction of candidate genes in *Magnaporthe oryzae* – a pilot study).

No	MGG ID	gene name	function	PHI-base ID
1	MGG_01818	MAGA	guanine nucleotide-binding protein alpha-3 subunit	PHI:82
2	MGG_04204	MAGC	guanine nucleotide-binding protein alpha-2 subunit	PHI:84

Appendix C.

Additional figures and tables to Chapter 5

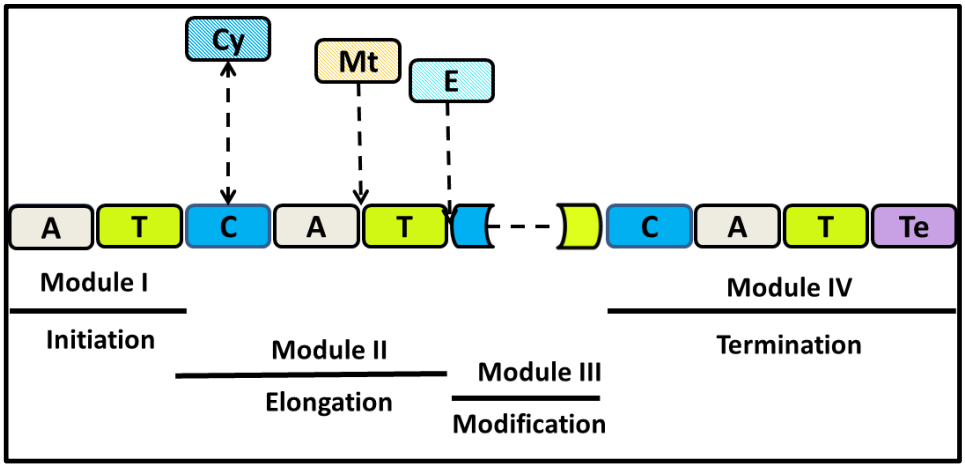


Figure C-1 Schematic modular organisation of NRPSs.
The order of modules and domains in complete NRPS is as follow: initiation module, elongation module, modification module and termination module. Adenylation (A) and thiolation or peptide carrier protein (T) domains are required by initiation and elongation modules. Elongation module in addition to A and T domains has condensation (C) domain. Hetero-cyclization (Cy) domain sometimes replaced C domain. Methyltransferase domain (Mt) and epimerization domain (E) are optional domains introducing modification module. Termination module usually has a thioesterase (Te) domain. This figure was generated based on the information provided by (Strieker et al., 2010).

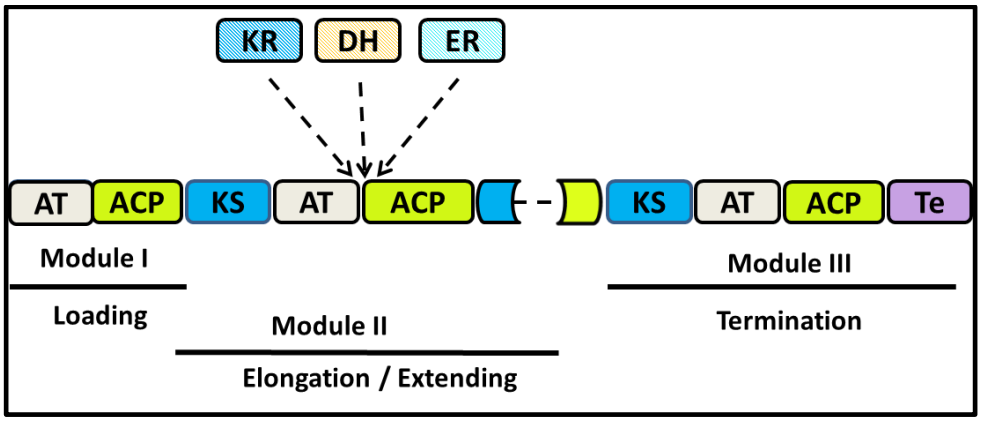


Figure C-2 Schematic modular organisation of PKSs.
Complete PKS consists of three main modules: loading module, elongation (extending) module and termination module. Acetyltransferase (AT) and acyl carrier protein (ACP) are required for loading and elongation (extending) modules. Elongation modules in addition to AT and ACP requires keto-synthase domain (KS). Ketoreductase domain (KR), dehydrtase domain (DH) and enoylreductase domain (ER) are optional domains introduced in elongation module between AT and ACP domains. Either one of those domains or all of them can be placed between AT and ACP domains. Termination module must have *thioesterase* domain. This figure was generated based on the information available in the study by Meier and Burkart (2009).

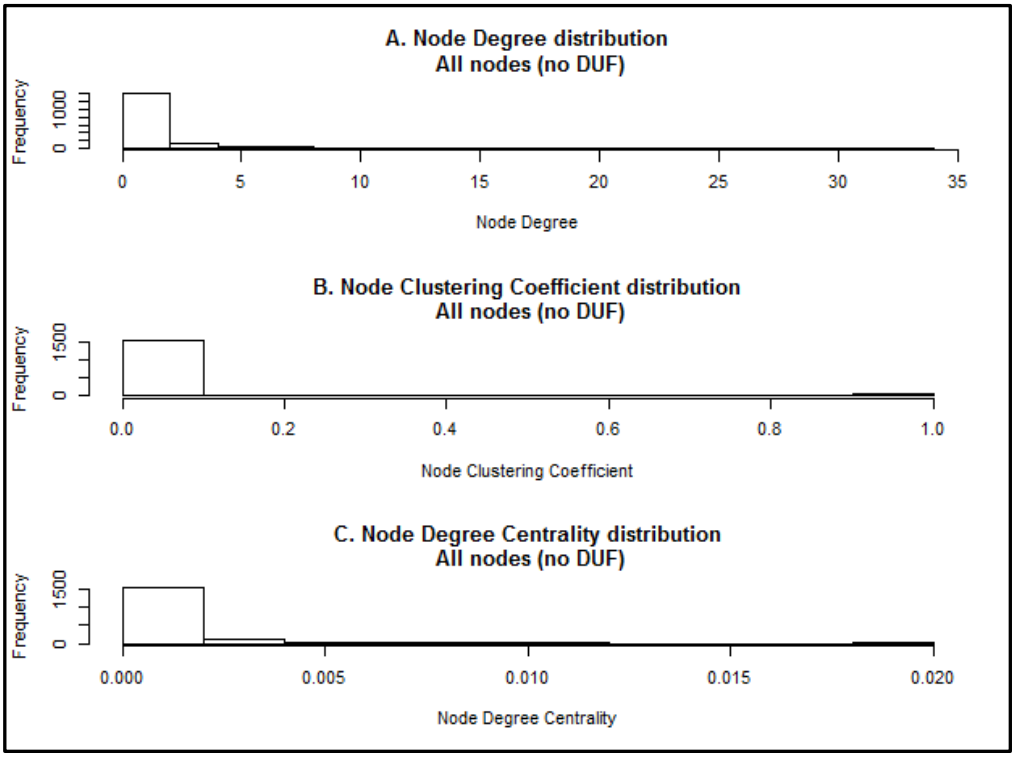


Figure C-3 Main topological properties of all nodes except for the DUFs nodes in the network

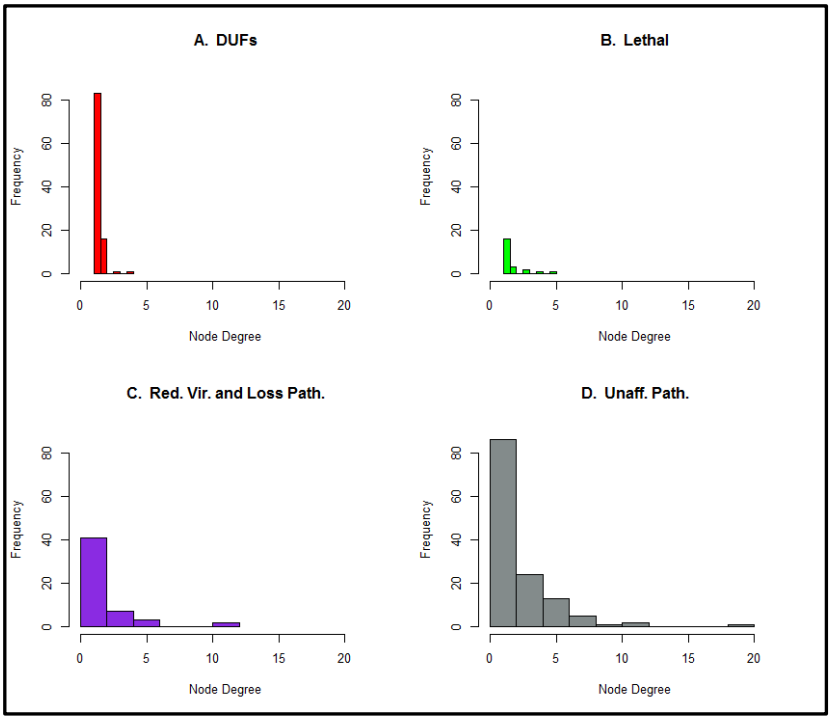


Figure C-4 Node degree distribution comparison within four sets of nodes.
A. DUFs nodes, B. Lethal nodes (not including DUFs), C. Reduced virulence and loss pathogenicity nodes (not including DUFs), and D. Unaffected pathogenicity nodes (not including DUFs).

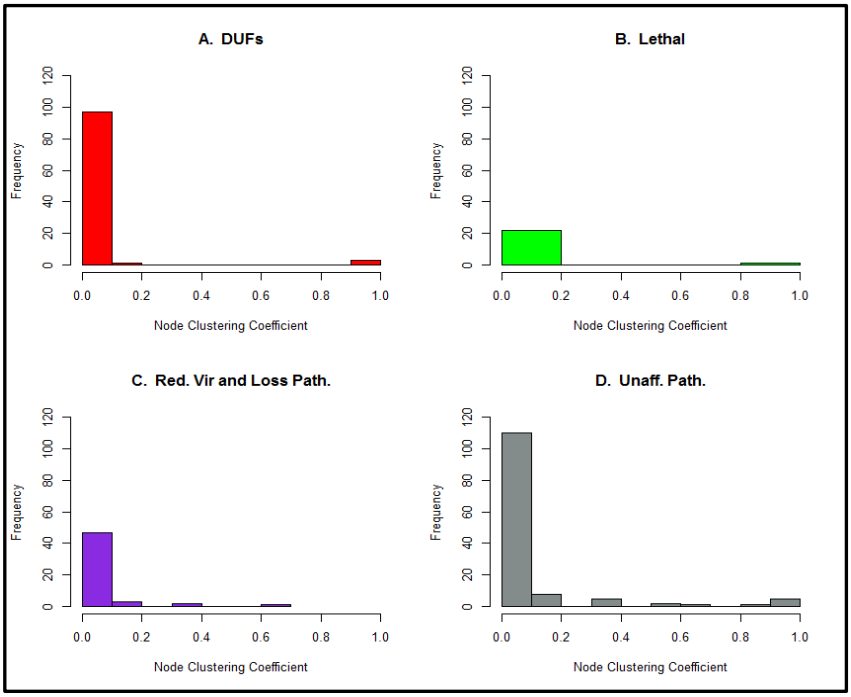


Figure C-5 Node clustering coefficient distribution comparison within four sets of nodes.
A. DUFs nodes, B. Lethal nodes (not including DUFs), C. Reduced virulence and loss pathogenicity nodes (not including DUFs), and D. Unaffected pathogenicity nodes (not including DUFs).

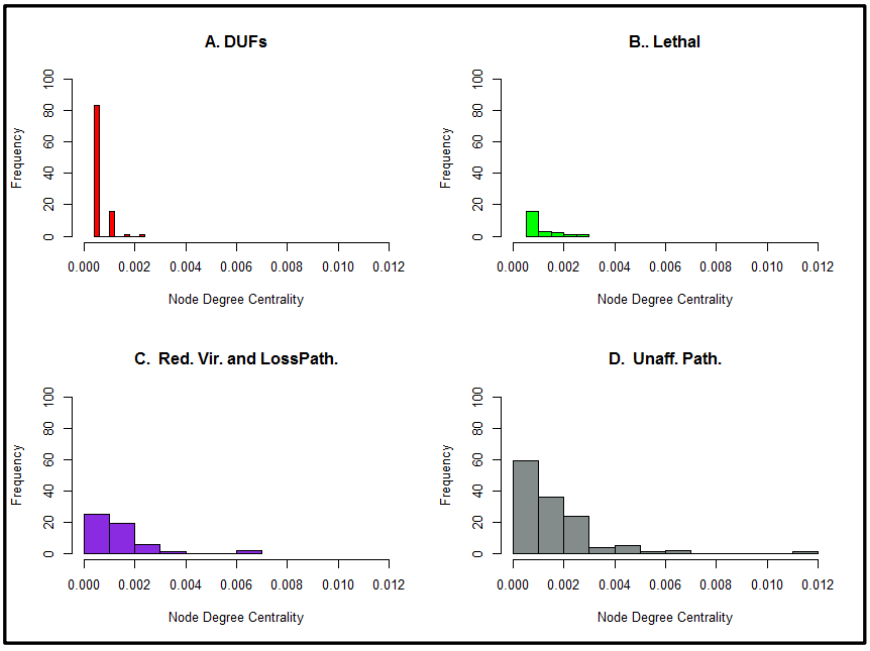


Figure C-6 Node degree centrality distribution comparison among four sets of nodes.
A. DUFs nodes, B. Lethal nodes (not including DUFs), C. Reduced virulence and loss pathogenicity nodes (not including DUFs), and D. Unaffected pathogenicity nodes (not including DUFs).

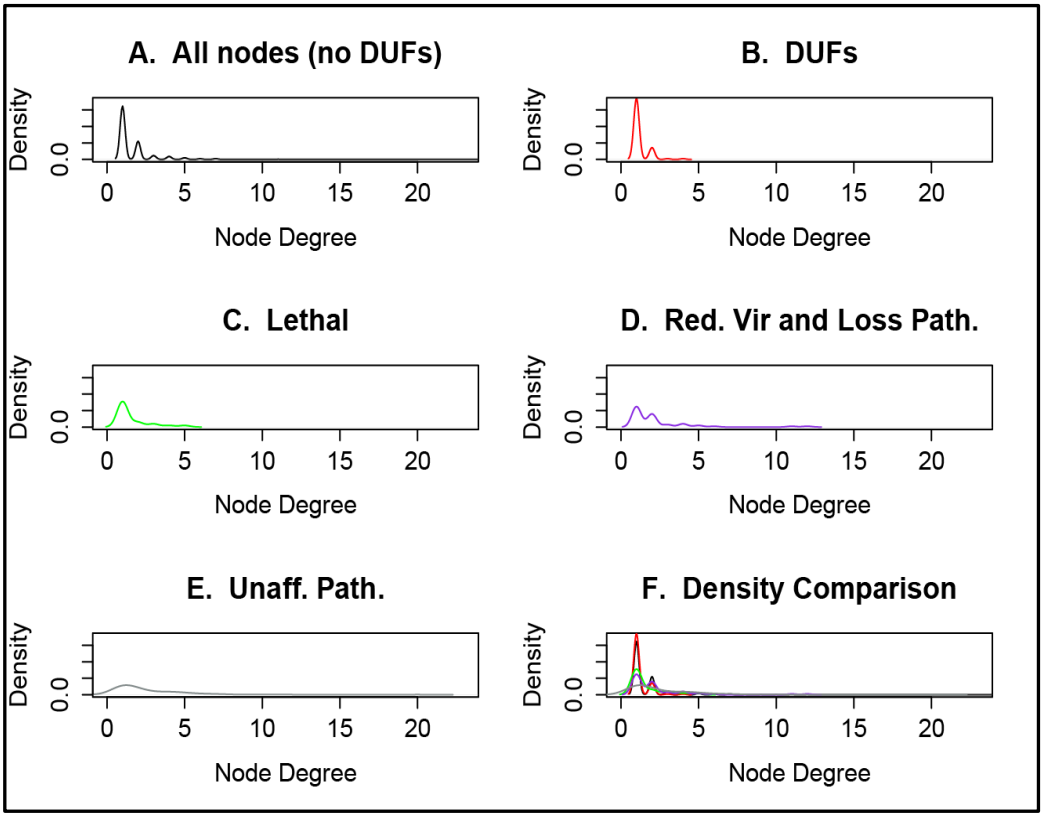


Figure C-7 Node degree Kernel Density Plots (KDPs) comparison.
A. All nodes except DUFs, B. DUF nodes, C. Lethal nodes (except DUFs), D. Reduced virulence and loss pathogenicity nodes (except DUFs), E. Unaffected pathogenicity nodes (except DUFs), F. Superimposition of KDPs: A., B., C., D., and E.

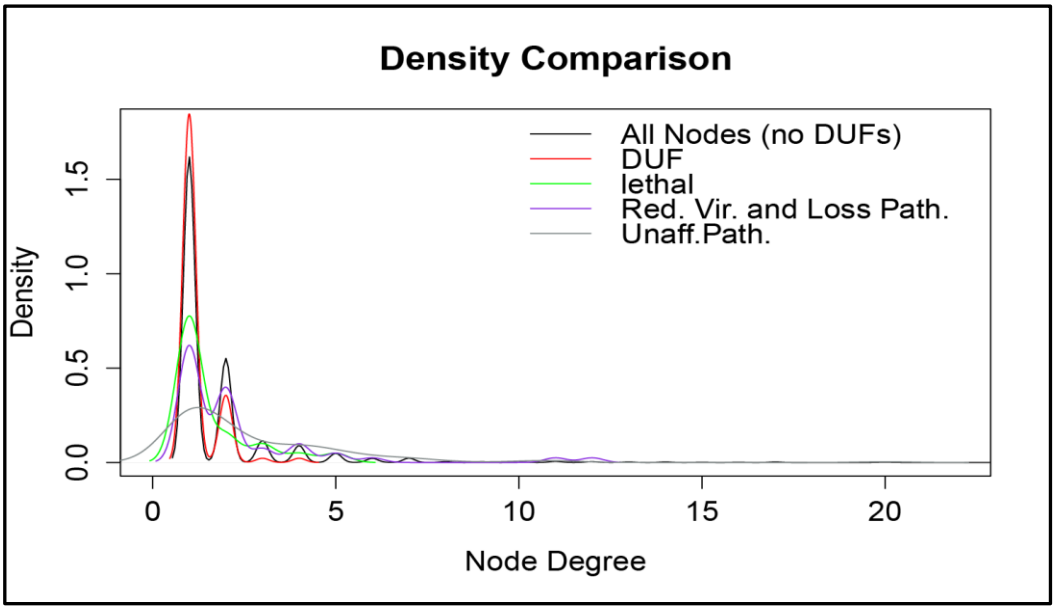


Figure C-8 Superimposition of node degree Kernel Density Plots.

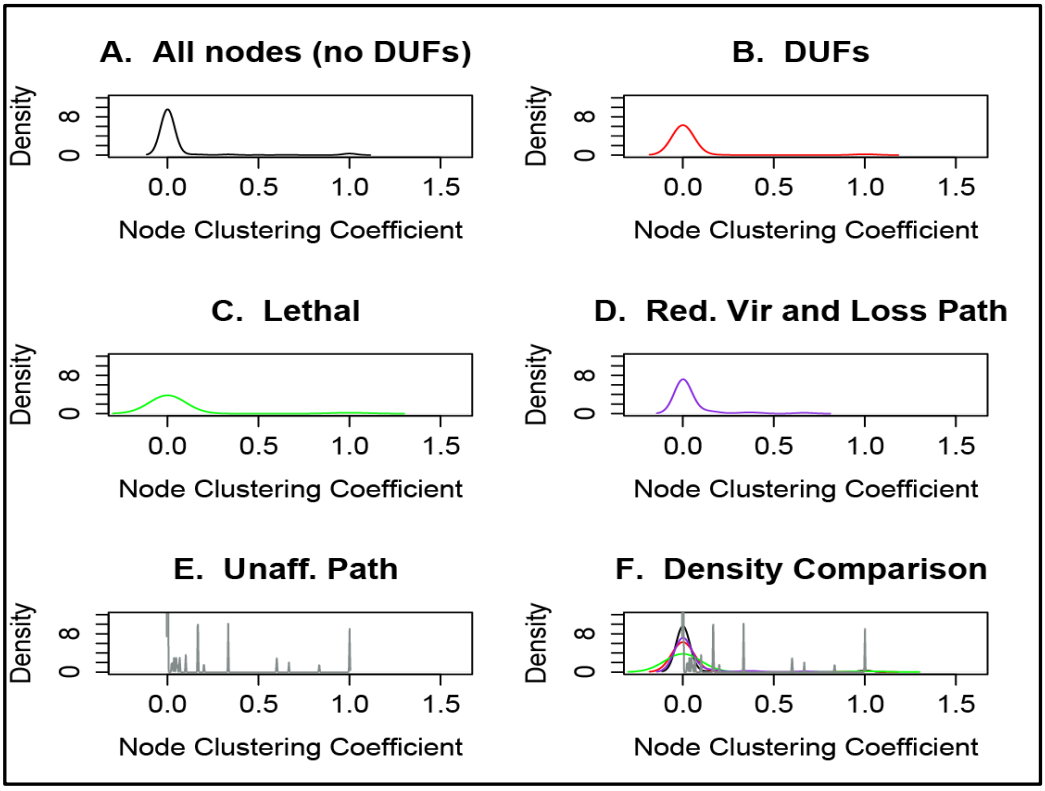


Figure C-9 Node clustering coefficient Kernel Density Plots (KDPs) comparison.
A. All nodes except DUFs, B. DUF nodes, C. Lethal nodes (except DUFs), D. Reduced virulence and loss pathogenicity nodes (except DUFs), E. Unaffected pathogenicity nodes (except DUFs), F. Superimposition of KDPs: A., B., C., D., and E.

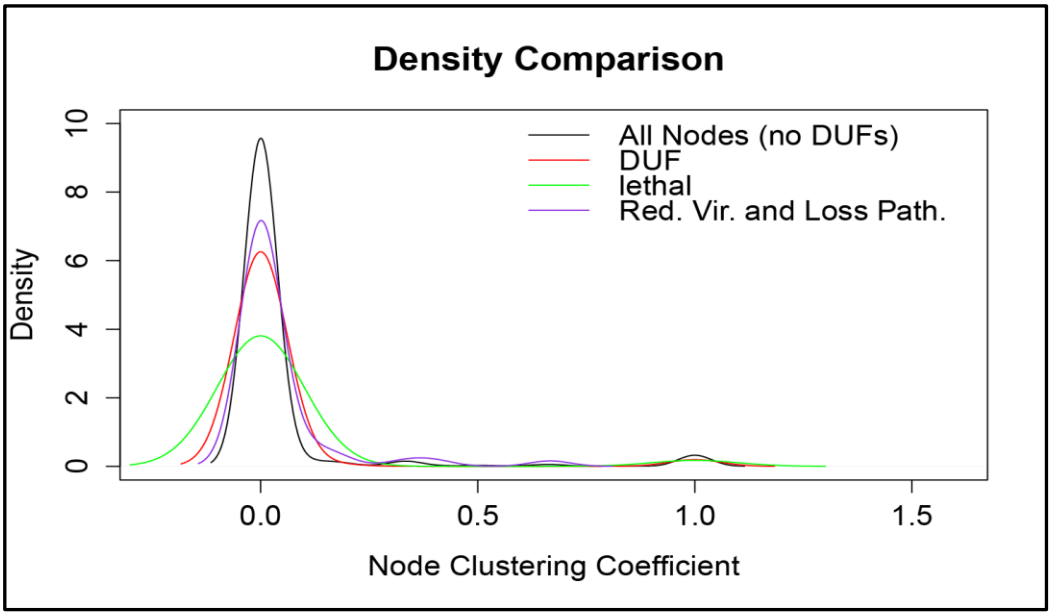


Figure C-10 Superimposition of node clustering coefficient Kernel Density Plots.

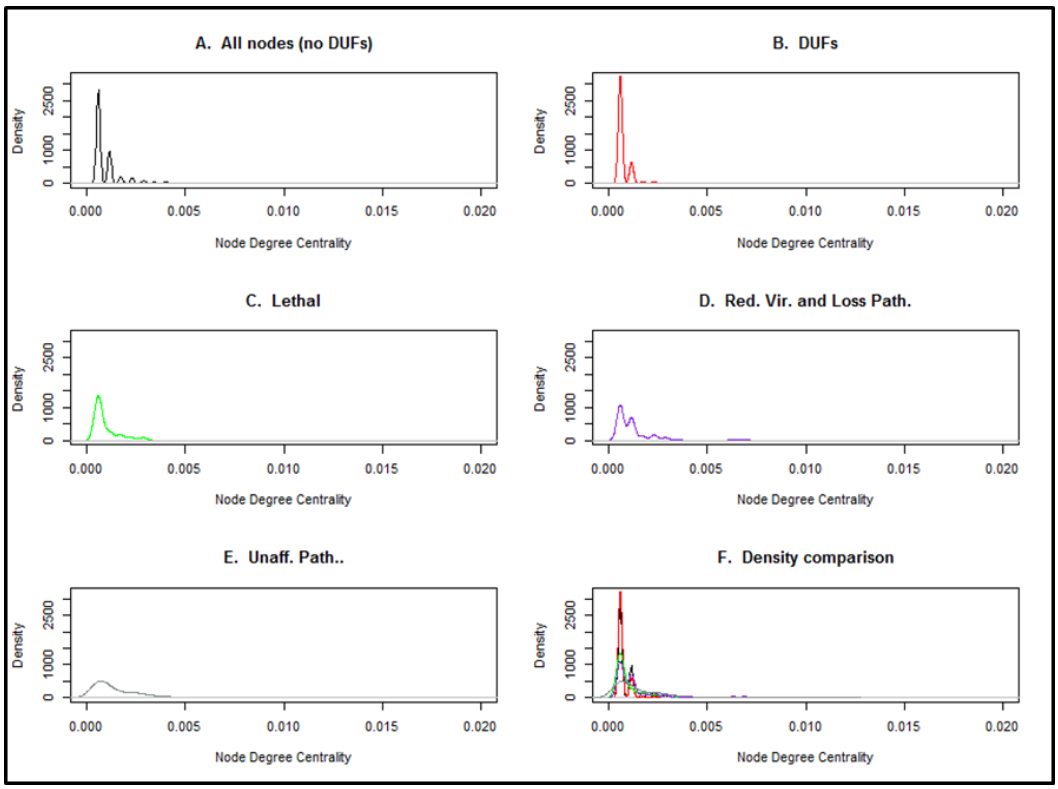


Figure C-11 Node degree centrality Kernel Density Plots (KDPs) comparison.
A. All nodes except DUFs, B. DUF nodes, C. Lethal nodes (except DUFs), D. Reduced virulence and loss pathogenicity nodes (except DUFs), E. Unaffected pathogenicity nodes (except DUFs), F. Superimposition of KDPs: A., B., C., D., and E.

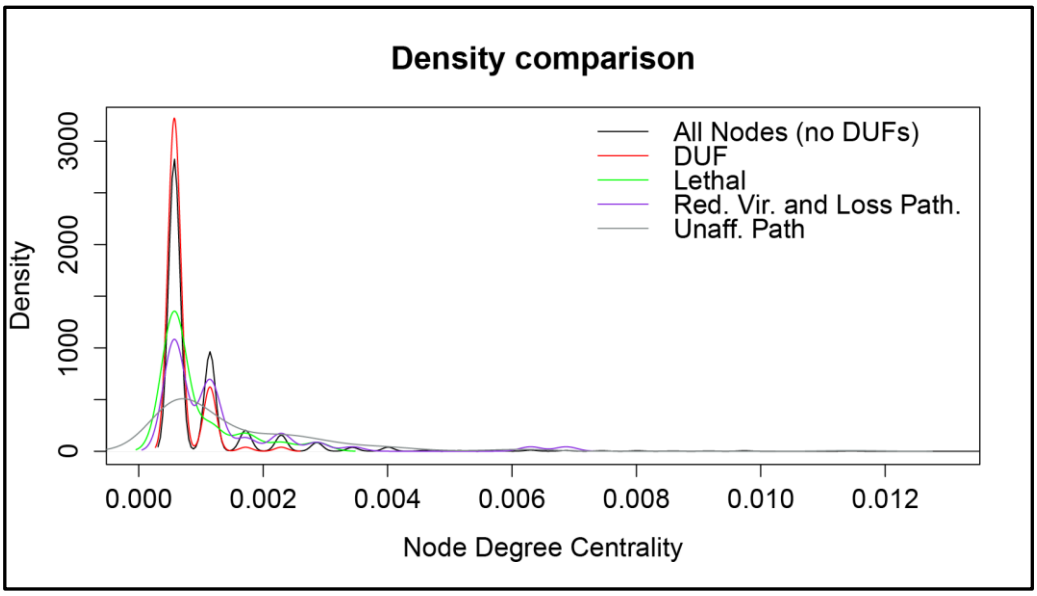


Figure C-12 Superimposition of node degree centrality Kernel Density Plots.

Table C-1 Frequency table calculated for the original fungi lifestyles categories.
This table is relevant to the text in Chapter 5, section 5.4.3.1 (Statistical evaluation of DUFs association with fungi lifestyle).

DUF Id	Life style	O	E	(O-E)2	(O-E)2/E	Life style	O	E*	(O-E)2	(O-E)2/E	Life style	O	E*	(O-E)2	(O-E)2/E	Life style	O	E	(O-E)2	(O-E)2/E	Life style	O	E	(O-E)2	(O-E)2/E
DUF2456	**PP**	23	20.01	8.92	0.45	**SP**	1	1.19	0.03	0.03	**FP**	3	1.54	2.14	1.39	**AP**	6	16.30	106.12	6.51	**NP**	28	21.96	36.47	1.66
DUF1965	**PP**	29	23.62	28.92	1.22	**SP**	2	1.40	0.36	0.26	**FP**	0	1.82	3.30	1.82	**AP**	14	19.24	27.47	1.43	**NP**	27	25.92	1.16	0.04
DUF3716	**PP**	29	24.93	16.53	0.66	**SP**	2	1.48	0.27	0.19	**FP**	2	1.92	0.01	0.00	**AP**	19	20.31	1.72	0.08	**NP**	24	27.36	11.30	0.41
DUF3129	**PP**	44	27.23	281.19	10.33	**SP**	4	1.61	5.70	3.53	**FP**	0	2.09	4.38	2.09	**AP**	16	22.18	38.20	1.72	**NP**	19	29.88	118.41	3.96
DUF2434	**PP**	26	27.56	2.43	0.09	**SP**	2	1.63	0.13	0.08	**FP**	3	2.12	0.78	0.37	**AP**	26	22.45	12.62	0.56	**NP**	27	30.24	10.51	0.35
DUF3176	**PP**	34	27.56	41.48	1.51	**SP**	2	1.63	0.13	0.08	**FP**	3	2.12	0.78	0.37	**AP**	18	22.45	19.78	0.88	**NP**	27	30.24	10.51	0.35
DUF3517	**PP**	35	31.50	12.27	0.39	**SP**	2	1.87	0.02	0.01	**FP**	3	2.42	0.34	0.14	**AP**	30	25.65	18.88	0.74	**NP**	26	34.56	73.30	2.12
DUF4045	**PP**	36	32.81	10.18	0.31	**SP**	2	1.94	0.00	0.00	**FP**	3	2.52	0.23	0.09	**AP**	30	26.72	10.73	0.40	**NP**	29	36.00	49.03	1.36
DUF4048	**PP**	39	33.14	34.38	1.04	**SP**	2	1.96	0.00	0.00	**FP**	3	2.55	0.21	0.08	**AP**	29	26.99	4.04	0.15	**NP**	28	36.36	69.92	1.92
DUF3636	**PP**	38	33.47	20.57	0.61	**SP**	1	1.98	0.97	0.49	**FP**	3	2.57	0.18	0.07	**AP**	30	27.26	7.52	0.28	**NP**	30	36.72	45.18	1.23
DUF3807	**PP**	39	33.79	27.11	0.80	**SP**	1	2.00	1.00	0.50	**FP**	3	2.60	0.16	0.06	**AP**	27	27.53	0.28	0.01	**NP**	33	37.08	16.66	0.45
DUF3984	**PP**	39	34.12	23.80	0.70	**SP**	2	2.02	0.00	0.00	**FP**	3	2.62	0.14	0.05	**AP**	31	27.79	10.29	0.37	**NP**	29	37.44	71.27	1.90
DUF2014	**PP**	41	35.43	30.98	0.87	**SP**	2	2.10	0.01	0.00	**FP**	3	2.72	0.08	0.03	**AP**	27	28.86	3.47	0.12	**NP**	35	38.88	15.07	0.39
DUF2457	**PP**	41	36.09	24.11	0.67	**SP**	1	2.14	1.29	0.61	**FP**	3	2.77	0.05	0.02	**AP**	32	29.40	6.78	0.23	**NP**	33	39.60	43.59	1.10
DUF3292	**PP**	47	36.42	111.98	3.07	**SP**	2	2.16	0.02	0.01	**FP**	3	2.80	0.04	0.01	**AP**	23	29.66	44.40	1.50	**NP**	36	39.96	15.70	0.39
DUF1774	**PP**	34	39.04	25.43	0.65	**SP**	2	2.31	0.10	0.04	**FP**	3	3.00	0.00	0.00	**AP**	38	31.80	38.42	1.21	**NP**	42	42.84	0.71	0.02
DUF3328	**PP**	43	39.04	15.66	0.40	**SP**	2	2.31	0.10	0.04	**FP**	3	3.00	0.00	0.00	**AP**	31	31.80	0.64	0.02	**NP**	40	42.84	8.08	0.19
DUF3433	**PP**	44	39.04	24.58	0.63	**SP**	3	2.31	0.47	0.20	**FP**	3	3.00	0.00	0.00	**AP**	31	31.80	0.64	0.02	**NP**	38	42.84	23.45	0.55
DUF4452	**PP**	36	39.37	11.36	0.29	**SP**	1	2.33	1.77	0.76	**FP**	3	3.03	0.00	0.00	**AP**	35	32.07	8.59	0.27	**NP**	45	43.20	3.23	0.07
DUF1770	**PP**	43	39.70	10.90	0.27	**SP**	2	2.35	0.12	0.05	**FP**	3	3.05	0.00	0.00	**AP**	32	32.34	0.11	0.00	**NP**	41	43.56	6.57	0.15
DUF3425	**PP**	48	43.64	19.05	0.44	**SP**	2	2.58	0.34	0.13	**FP**	3	3.35	0.13	0.04	**AP**	36	35.54	0.21	0.01	**NP**	44	47.88	15.07	0.31
DUF4484	**PP**	41	45.60	21.20	0.46	**SP**	2	2.70	0.49	0.18	**FP**	4	3.51	0.24	0.07	**AP**	39	37.15	3.44	0.09	**NP**	53	50.04	8.75	0.17
DUF2011	**PP**	38	46.92	79.51	1.69	**SP**	1	2.78	3.17	1.14	**FP**	4	3.61	0.15	0.04	**AP**	42	38.21	14.33	0.37	**NP**	58	51.48	42.48	0.83
DUF3812	**PP**	40	48.23	67.72	1.40	**SP**	2	2.86	0.73	0.26	**FP**	4	3.71	0.09	0.02	**AP**	40	39.28	0.51	0.01	**NP**	61	52.92	65.24	1.23
DUF2406	**PP**	41	48.56	57.11	1.18	**SP**	2	2.88	0.77	0.27	**FP**	4	3.73	0.07	0.02	**AP**	38	39.55	2.41	0.06	**NP**	63	53.28	94.43	1.77
DUF1691	**PP**	44	48.56	20.77	0.43	**SP**	4	2.88	1.26	0.44	**FP**	4	3.73	0.07	0.02	**AP**	41	39.55	2.10	0.05	**NP**	55	53.28	2.95	0.06
DUF4448	**PP**	39	49.21	104.31	2.12	**SP**	4	2.92	1.18	0.40	**FP**	3	3.78	0.61	0.16	**AP**	39	40.09	1.18	0.03	**NP**	65	54.00	120.94	2.24
DUF3115	**PP**	42	49.54	56.87	1.15	**SP**	2	2.93	0.87	0.30	**FP**	4	3.81	0.04	0.01	**AP**	44	40.35	13.30	0.33	**NP**	59	54.36	21.50	0.40
DUF2417	**PP**	47	50.85	14.85	0.29	**SP**	2	3.01	1.03	0.34	**FP**	4	3.91	0.01	0.00	**AP**	40	41.42	2.02	0.05	**NP**	62	55.80	38.40	0.69
DUF4451	**PP**	48	52.17	17.36	0.33	**SP**	4	3.09	0.83	0.27	**FP**	3	4.01	1.02	0.25	**AP**	46	42.49	12.31	0.29	**NP**	58	57.24	0.57	0.01
DUF1687	**PP**	45	52.49	56.16	1.07	**SP**	5	3.11	3.57	1.15	**FP**	4	4.04	0.00	0.00	**AP**	39	42.76	14.12	0.33	**NP**	67	57.60	88.30	1.53
DUF3844	**PP**	51	53.15	4.62	0.09	**SP**	4	3.15	0.72	0.23	**FP**	3	4.09	1.18	0.29	**AP**	48	43.29	22.16	0.51	**NP**	56	58.32	5.40	0.09
DUF3835	**PP**	47	54.13	50.90	0.94	**SP**	5	3.21	3.22	1.00	**FP**	4	4.16	0.03	0.01	**AP**	43	44.09	1.20	0.03	**NP**	66	59.40	43.52	0.73
DUF3602	**PP**	53	59.38	40.76	0.69	**SP**	4	3.52	0.23	0.07	**FP**	4	4.56	0.32	0.07	**AP**	45	48.37	11.36	0.23	**NP**	75	65.16	96.76	1.48
DUF3779	**PP**	54	61.68	58.99	0.96	**SP**	3	3.65	0.43	0.12	**FP**	4	4.74	0.55	0.12	**AP**	50	50.24	0.06	0.00	**NP**	77	67.68	86.80	1.28

PP - plant pathogenic fungi, SP –symbionts of plant roots and endophyte, FP – pathogens of fungi (fungi infecting other fungi), AP – animal pathogens (fungi infecting animals), NP - non-pathogenic fungi, O- observed frequencies, E – expected frequencies. *Highlighted frequencies with value less than 5.

Table C-2 Frequency table for chi-square test 1.
This table is relevant to the text in Chapter 5, section 5.4.3.1 (Statistical evaluation of DUFs association with fungi lifestyle).

DUF Id	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E
DUF2456	**PP**	23	20.01	2.99	8.92	0.45	**SPFP**	4	2.72	1.28	1.63	0.60	**AP**	6	16.30	-10.30	106.12	6.51	**NP**	28	21.96	6.04	36.47	1.66
DUF1965	**PP**	29	23.62	5.38	28.92	1.22	**SPFP**	2	3.22	-1.22	1.48	0.46	**AP**	14	19.24	-5.24	27.47	1.43	**NP**	27	25.92	1.08	1.16	0.04
DUF3716	**PP**	29	24.93	4.07	16.53	0.66	**SPFP**	4	3.39	0.61	0.37	0.11	**AP**	19	20.31	-1.31	1.72	0.08	**NP**	24	27.36	-3.36	11.30	0.41
DUF3129	**PP**	44	27.23	16.77	281.19	10.33	**SPFP**	4	3.71	0.29	0.09	0.02	**AP**	16	22.18	-6.18	38.20	1.72	**NP**	19	29.88	-10.88	118.41	3.96
DUF2434	**PP**	26	27.56	-1.56	2.43	0.09	**SPFP**	5	3.75	1.25	1.56	0.42	**AP**	26	22.45	3.55	12.62	0.56	**NP**	27	30.24	-3.24	10.51	0.35
DUF3176	**PP**	34	27.56	6.44	41.48	1.51	**SPFP**	5	3.75	1.25	1.56	0.42	**AP**	18	22.45	-4.45	19.78	0.88	**NP**	27	30.24	-3.24	10.51	0.35
DUF3517	**PP**	35	31.50	3.50	12.27	0.39	**SPFP**	5	4.29	0.71	0.51	0.12	**AP**	30	25.65	4.35	18.88	0.74	**NP**	26	34.56	-8.56	73.30	2.12
DUF4045	**PP**	36	32.81	3.19	10.18	0.31	**SPFP**	5	4.47	0.53	0.29	0.06	**AP**	30	26.72	3.28	10.73	0.40	**NP**	29	36.00	-7.00	49.03	1.36
DUF4048	**PP**	39	33.14	5.86	34.38	1.04	**SPFP**	5	4.51	0.49	0.24	0.05	**AP**	29	26.99	2.01	4.04	0.15	**NP**	28	36.36	-8.36	69.92	1.92
DUF3636	**PP**	38	33.47	4.53	20.57	0.61	**SPFP**	4	4.55	-0.55	0.31	0.07	**AP**	30	27.26	2.74	7.52	0.28	**NP**	30	36.72	-6.72	45.18	1.23
DUF3807	**PP**	39	33.79	5.21	27.11	0.80	**SPFP**	4	4.60	-0.60	0.36	0.08	**AP**	27	27.53	-0.53	0.28	0.01	**NP**	33	37.08	-4.08	16.66	0.45
DUF3984	**PP**	39	34.12	4.88	23.80	0.70	**SPFP**	5	4.64	0.36	0.13	0.03	**AP**	31	27.79	3.21	10.29	0.37	**NP**	29	37.44	-8.44	71.27	1.90
DUF2014	**PP**	41	35.43	5.57	30.98	0.87	**SPFP**	5	4.82	0.18	0.03	0.01	**AP**	27	28.86	-1.86	3.47	0.12	**NP**	35	38.88	-3.88	15.07	0.39
DUF2457	**PP**	41	36.09	4.91	24.11	0.67	**SPFP**	4	4.91	-0.91	0.83	0.17	**AP**	32	29.40	2.60	6.78	0.23	**NP**	33	39.60	-6.60	43.59	1.10
DUF3292	**PP**	47	36.42	10.58	111.98	3.07	**SPFP**	5	4.96	0.04	0.00	0.00	**AP**	23	29.66	-6.66	44.40	1.50	**NP**	36	39.96	-3.96	15.70	0.39
DUF1774	**PP**	34	39.04	-5.04	25.43	0.65	**SPFP**	5	5.31	-0.31	0.10	0.02	**AP**	38	31.80	6.20	38.42	1.21	**NP**	42	42.84	-0.84	0.71	0.02
DUF3328	**PP**	43	39.04	3.96	15.66	0.40	**SPFP**	5	5.31	-0.31	0.10	0.02	**AP**	31	31.80	-0.80	0.64	0.02	**NP**	40	42.84	-2.84	8.08	0.19
DUF3433	**PP**	44	39.04	4.96	24.58	0.63	**SPFP**	6	5.31	0.69	0.47	0.09	**AP**	31	31.80	-0.80	0.64	0.02	**NP**	38	42.84	-4.84	23.45	0.55
DUF4452	**PP**	36	39.37	-3.37	11.36	0.29	**SPFP**	4	5.36	-1.36	1.85	0.34	**AP**	35	32.07	2.93	8.59	0.27	**NP**	45	43.20	1.80	3.23	0.07
DUF1770	**PP**	43	39.70	3.30	10.90	0.27	**SPFP**	5	5.40	-0.40	0.16	0.03	**AP**	32	32.34	-0.34	0.11	0.00	**NP**	41	43.56	-2.56	6.57	0.15
DUF3425	**PP**	48	43.64	4.36	19.05	0.44	**SPFP**	5	5.94	-0.94	0.88	0.15	**AP**	36	35.54	0.46	0.21	0.01	**NP**	44	47.88	-3.88	15.07	0.31
DUF4484	**PP**	41	45.60	-4.60	21.20	0.46	**SPFP**	6	6.21	-0.21	0.04	0.01	**AP**	39	37.15	1.85	3.44	0.09	**NP**	53	50.04	2.96	8.75	0.17
DUF2011	**PP**	38	46.92	-8.92	79.51	1.69	**SPFP**	5	6.39	-1.39	1.92	0.30	**AP**	42	38.21	3.79	14.33	0.37	**NP**	58	51.48	6.52	42.48	0.83
DUF3812	**PP**	40	48.23	-8.23	67.72	1.40	**SPFP**	6	6.56	-0.56	0.32	0.05	**AP**	40	39.28	0.72	0.51	0.01	**NP**	61	52.92	8.08	65.24	1.23
DUF2406	**PP**	41	48.56	-7.56	57.11	1.18	**SPFP**	6	6.61	-0.61	0.37	0.06	**AP**	38	39.55	-1.55	2.41	0.06	**NP**	63	53.28	9.72	94.43	1.77
DUF1691	**PP**	44	48.56	-4.56	20.77	0.43	**SPFP**	8	6.61	1.39	1.93	0.29	**AP**	41	39.55	1.45	2.10	0.05	**NP**	55	53.28	1.72	2.95	0.06
DUF4448	**PP**	39	49.21	-10.21	104.31	2.12	**SPFP**	7	6.70	0.30	0.09	0.01	**AP**	39	40.09	-1.09	1.18	0.03	**NP**	65	54.00	11.00	120.94	2.24
DUF3115	**PP**	42	49.54	-7.54	56.87	1.15	**SPFP**	6	6.74	-0.74	0.55	0.08	**AP**	44	40.35	3.65	13.30	0.33	**NP**	59	54.36	4.64	21.50	0.40
DUF2417	**PP**	47	50.85	-3.85	14.85	0.29	**SPFP**	6	6.92	-0.92	0.85	0.12	**AP**	40	41.42	-1.42	2.02	0.05	**NP**	62	55.80	6.20	38.40	0.69
DUF4451	**PP**	48	52.17	-4.17	17.36	0.33	**SPFP**	7	7.10	-0.10	0.01	0.00	**AP**	46	42.49	3.51	12.31	0.29	**NP**	58	57.24	0.76	0.57	0.01
DUF1687	**PP**	45	52.49	-7.49	56.16	1.07	**SPFP**	9	7.14	1.86	3.44	0.48	**AP**	39	42.76	-3.76	14.12	0.33	**NP**	67	57.60	9.40	88.30	1.53
DUF3844	**PP**	51	53.15	-2.15	4.62	0.09	**SPFP**	7	7.23	-0.23	0.05	0.01	**AP**	48	43.29	4.71	22.16	0.51	**NP**	56	58.32	-2.32	5.40	0.09
DUF3835	**PP**	47	54.13	-7.13	50.90	0.94	**SPFP**	9	7.37	1.63	2.66	0.36	**AP**	43	44.09	-1.09	1.20	0.03	**NP**	66	59.40	6.60	43.52	0.73
DUF3602	**PP**	53	59.38	-6.38	40.76	0.69	**SPFP**	8	8.08	-0.08	0.01	0.00	**AP**	45	48.37	-3.37	11.36	0.23	**NP**	75	65.16	9.84	96.76	1.48
DUF3779	**PP**	54	61.68	-7.68	58.99	0.96	**SPFP**	7	8.40	-1.40	1.95	0.23	**AP**	50	50.24	-0.24	0.06	0.00	**NP**	77	67.68	9.32	86.80	1.28

PP - plant pathogenic fungi, SPFP – includes lifestyle groups: symbionts of plant roots and endophyte (SP) and fungi infecting other fungi (FP), AP – fungi infecting animals, NP - non-pathogenic fungi, O- observed frequencies, E – expected frequencies.

Table C-3: Frequency table for chi-square test 2.
This table is relevant to the text in Chapter 5, section 5.4.3.1 (Statistical evaluation of DUFs association with fungi lifestyle).

DUF Id	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E
DUF2456	**PP**	23	20.01	2.99	8.92	0.45	**OP**	10	19.03	-9.03	81.46	4.28	**NP**	28	21.96	6.04	36.47	1.66
DUF1965	**PP**	29	23.62	5.38	28.92	1.22	**OP**	16	22.46	-6.46	41.68	1.86	**NP**	27	25.92	1.08	1.16	0.04
DUF3716	**PP**	29	24.93	4.07	16.53	0.66	**OP**	23	23.70	-0.70	0.50	0.02	**NP**	24	27.36	-3.36	11.30	0.41
DUF3129	**PP**	44	27.23	16.77	281.19	10.33	**OP**	20	25.89	-5.89	34.66	1.34	**NP**	19	29.88	-10.88	118.41	3.96
DUF2434	**PP**	26	27.56	-1.56	2.43	0.09	**OP**	31	26.20	4.80	23.05	0.88	**NP**	27	30.24	-3.24	10.51	0.35
DUF3176	**PP**	34	27.56	6.44	41.48	1.51	**OP**	23	26.20	-3.20	10.23	0.39	**NP**	27	30.24	-3.24	10.51	0.35
DUF3517	**PP**	35	31.50	3.50	12.27	0.39	**OP**	35	29.94	5.06	25.59	0.85	**NP**	26	34.56	-8.56	73.30	2.12
DUF4045	**PP**	36	32.81	3.19	10.18	0.31	**OP**	35	31.19	3.81	14.52	0.47	**NP**	29	36.00	-7.00	49.03	1.36
DUF4048	**PP**	39	33.14	5.86	34.38	1.04	**OP**	34	31.50	2.50	6.24	0.20	**NP**	28	36.36	-8.36	69.92	1.92
DUF3636	**PP**	38	33.47	4.53	20.57	0.61	**OP**	34	31.81	2.19	4.78	0.15	**NP**	30	36.72	-6.72	45.18	1.23
DUF3807	**PP**	39	33.79	5.21	27.11	0.80	**OP**	31	32.12	-1.12	1.27	0.04	**NP**	33	37.08	-4.08	16.66	0.45
DUF3984	**PP**	39	34.12	4.88	23.80	0.70	**OP**	36	32.44	3.56	12.70	0.39	**NP**	29	37.44	-8.44	71.27	1.90
DUF2014	**PP**	41	35.43	5.57	30.98	0.87	**OP**	32	33.68	-1.68	2.84	0.08	**NP**	35	38.88	-3.88	15.07	0.39
DUF2457	**PP**	41	36.09	4.91	24.11	0.67	**OP**	36	34.31	1.69	2.86	0.08	**NP**	33	39.60	-6.60	43.59	1.10
DUF3292	**PP**	47	36.42	10.58	111.98	3.07	**OP**	28	34.62	-6.62	43.83	1.27	**NP**	36	39.96	-3.96	15.70	0.39
DUF1774	**PP**	34	39.04	-5.04	25.43	0.65	**OP**	43	37.12	5.88	34.63	0.93	**NP**	42	42.84	-0.84	0.71	0.02
DUF3328	**PP**	43	39.04	3.96	15.66	0.40	**OP**	36	37.12	-1.12	1.24	0.03	**NP**	40	42.84	-2.84	8.08	0.19
DUF3433	**PP**	44	39.04	4.96	24.58	0.63	**OP**	37	37.12	-0.12	0.01	0.00	**NP**	38	42.84	-4.84	23.45	0.55
DUF4452	**PP**	36	39.37	-3.37	11.36	0.29	**OP**	39	37.43	1.57	2.47	0.07	**NP**	45	43.20	1.80	3.23	0.07
DUF1770	**PP**	43	39.70	3.30	10.90	0.27	**OP**	37	37.74	-0.74	0.55	0.01	**NP**	41	43.56	-2.56	6.57	0.15
DUF3425	**PP**	48	43.64	4.36	19.05	0.44	**OP**	41	41.48	-0.48	0.23	0.01	**NP**	44	47.88	-3.88	15.07	0.31
DUF4484	**PP**	41	45.60	-4.60	21.20	0.46	**OP**	45	43.35	1.65	2.71	0.06	**NP**	53	50.04	2.96	8.75	0.17
DUF2011	**PP**	38	46.92	-8.92	79.51	1.69	**OP**	47	44.60	2.40	5.76	0.13	**NP**	58	51.48	6.52	42.48	0.83
DUF3812	**PP**	40	48.23	-8.23	67.72	1.40	**OP**	46	45.85	0.15	0.02	0.00	**NP**	61	52.92	8.08	65.24	1.23
DUF2406	**PP**	41	48.56	-7.56	57.11	1.18	**OP**	44	46.16	-2.16	4.67	0.10	**NP**	63	53.28	9.72	94.43	1.77
DUF1691	**PP**	44	48.56	-4.56	20.77	0.43	**OP**	49	46.16	2.84	8.06	0.17	**NP**	55	53.28	1.72	2.95	0.06
DUF4448	**PP**	39	49.21	-10.21	104.31	2.12	**OP**	46	46.78	-0.78	0.61	0.01	**NP**	65	54.00	11.00	120.94	2.24
DUF3115	**PP**	42	49.54	-7.54	56.87	1.15	**OP**	50	47.10	2.90	8.43	0.18	**NP**	59	54.36	4.64	21.50	0.40
DUF2417	**PP**	47	50.85	-3.85	14.85	0.29	**OP**	46	48.34	-2.34	5.49	0.11	**NP**	62	55.80	6.20	38.40	0.69
DUF4451	**PP**	48	52.17	-4.17	17.36	0.33	**OP**	53	49.59	3.41	11.62	0.23	**NP**	58	57.24	0.76	0.57	0.01
DUF1687	**PP**	45	52.49	-7.49	56.16	1.07	**OP**	48	49.90	-1.90	3.62	0.07	**NP**	67	57.60	9.40	88.30	1.53
DUF3844	**PP**	51	53.15	-2.15	4.62	0.09	**OP**	55	50.53	4.47	20.01	0.40	**NP**	56	58.32	-2.32	5.40	0.09
DUF3835	**PP**	47	54.13	-7.13	50.90	0.94	**OP**	52	51.46	0.54	0.29	0.01	**NP**	66	59.40	6.60	43.52	0.73
DUF3602	**PP**	53	59.38	-6.38	40.76	0.69	**OP**	53	56.45	-3.45	11.92	0.21	**NP**	75	65.16	9.84	96.76	1.48
DUF3779	**PP**	54	61.68	-7.68	58.99	0.96	**OP**	57	58.64	-1.64	2.68	0.05	**NP**	77	67.68	9.32	86.80	1.28

PP - plant pathogens, OP – other pathogens (includes lifestyle groups: symbionts of plant roots and endophyte (SP), fungi infecting other fungi (FP) and AP – fungi infecting animals), NP - non-pathogenic fungi, O- observed frequencies, E – expected frequencies.

Table C-4: Frequency table for chi-square test 3.
This table is relevant to the text in Chapter 5, section 5.4.3.1 (Statistical evaluation of DUFs association with fungi lifestyle).

DUF Id	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E
DUF2456	**PP+SPFP**	27	22.74	4.26	18.17	0.80	**AP**	6	16.30	-10.30	106.12	6.51	**NP**	28	21.96	6.04	36.47	1.66
DUF1965	**PP+SPFP**	31	26.84	4.16	17.33	0.65	**AP**	14	19.24	-5.24	27.47	1.43	**NP**	27	25.92	1.08	1.16	0.04
DUF3716	**PP+SPFP**	33	28.33	4.67	21.82	0.77	**AP**	19	20.31	-1.31	1.72	0.08	**NP**	24	27.36	-3.36	11.30	0.41
DUF3129	**PP+SPFP**	48	30.94	17.06	291.12	9.41	**AP**	16	22.18	-6.18	38.20	1.72	**NP**	19	29.88	-10.88	118.41	3.96
DUF2434	**PP+SPFP**	31	31.31	-0.31	0.10	0.00	**AP**	26	22.45	3.55	12.62	0.56	**NP**	27	30.24	-3.24	10.51	0.35
DUF3176	**PP+SPFP**	39	31.31	7.69	59.13	1.89	**AP**	18	22.45	-4.45	19.78	0.88	**NP**	27	30.24	-3.24	10.51	0.35
DUF3517	**PP+SPFP**	40	35.78	4.22	17.78	0.50	**AP**	30	25.65	4.35	18.88	0.74	**NP**	26	34.56	-8.56	73.30	2.12
DUF4045	**PP+SPFP**	41	37.27	3.73	13.88	0.37	**AP**	30	26.72	3.28	10.73	0.40	**NP**	29	36.00	-7.00	49.03	1.36
DUF4048	**PP+SPFP**	44	37.65	6.35	40.36	1.07	**AP**	29	26.99	2.01	4.04	0.15	**NP**	28	36.36	-8.36	69.92	1.92
DUF3636	**PP+SPFP**	42	38.02	3.98	15.84	0.42	**AP**	30	27.26	2.74	7.52	0.28	**NP**	30	36.72	-6.72	45.18	1.23
DUF3807	**PP+SPFP**	43	38.39	4.61	21.23	0.55	**AP**	27	27.53	-0.53	0.28	0.01	**NP**	33	37.08	-4.08	16.66	0.45
DUF3984	**PP+SPFP**	44	38.77	5.23	27.40	0.71	**AP**	31	27.79	3.21	10.29	0.37	**NP**	29	37.44	-8.44	71.27	1.90
DUF2014	**PP+SPFP**	46	40.26	5.74	32.99	0.82	**AP**	27	28.86	-1.86	3.47	0.12	**NP**	35	38.88	-3.88	15.07	0.39
DUF2457	**PP+SPFP**	45	41.00	4.00	15.99	0.39	**AP**	32	29.40	2.60	6.78	0.23	**NP**	33	39.60	-6.60	43.59	1.10
DUF3292	**PP+SPFP**	52	41.37	10.63	112.90	2.73	**AP**	23	29.66	-6.66	44.40	1.50	**NP**	36	39.96	-3.96	15.70	0.39
DUF1774	**PP+SPFP**	39	44.36	-5.36	28.69	0.65	**AP**	38	31.80	6.20	38.42	1.21	**NP**	42	42.84	-0.84	0.71	0.02
DUF3328	**PP+SPFP**	48	44.36	3.64	13.27	0.30	**AP**	31	31.80	-0.80	0.64	0.02	**NP**	40	42.84	-2.84	8.08	0.19
DUF3433	**PP+SPFP**	50	44.36	5.64	31.85	0.72	**AP**	31	31.80	-0.80	0.64	0.02	**NP**	38	42.84	-4.84	23.45	0.55
DUF4452	**PP+SPFP**	40	44.73	-4.73	22.37	0.50	**AP**	35	32.07	2.93	8.59	0.27	**NP**	45	43.20	1.80	3.23	0.07
DUF1770	**PP+SPFP**	48	45.10	2.90	8.40	0.19	**AP**	32	32.34	-0.34	0.11	0.00	**NP**	41	43.56	-2.56	6.57	0.15
DUF3425	**PP+SPFP**	53	49.57	3.43	11.73	0.24	**AP**	36	35.54	0.46	0.21	0.01	**NP**	44	47.88	-3.88	15.07	0.31
DUF4484	**PP+SPFP**	47	51.81	-4.81	23.15	0.45	**AP**	39	37.15	1.85	3.44	0.09	**NP**	53	50.04	2.96	8.75	0.17
DUF2011	**PP+SPFP**	43	53.30	-10.30	106.14	1.99	**AP**	42	38.21	3.79	14.33	0.37	**NP**	58	51.48	6.52	42.48	0.83
DUF3812	**PP+SPFP**	46	54.79	-8.79	77.32	1.41	**AP**	40	39.28	0.72	0.51	0.01	**NP**	61	52.92	8.08	65.24	1.23
DUF2406	**PP+SPFP**	47	55.17	-8.17	66.69	1.21	**AP**	38	39.55	-1.55	2.41	0.06	**NP**	63	53.28	9.72	94.43	1.77
DUF1691	**PP+SPFP**	52	55.17	-3.17	10.02	0.18	**AP**	41	39.55	1.45	2.10	0.05	**NP**	55	53.28	1.72	2.95	0.06
DUF4448	**PP+SPFP**	46	55.91	-9.91	98.24	1.76	**AP**	39	40.09	-1.09	1.18	0.03	**NP**	65	54.00	11.00	120.94	2.24
DUF3115	**PP+SPFP**	48	56.28	-8.28	68.63	1.22	**AP**	44	40.35	3.65	13.30	0.33	**NP**	59	54.36	4.64	21.50	0.40
DUF2417	**PP+SPFP**	53	57.78	-4.78	22.80	0.39	**AP**	40	41.42	-1.42	2.02	0.05	**NP**	62	55.80	6.20	38.40	0.69
DUF4451	**PP+SPFP**	55	59.27	-4.27	18.20	0.31	**AP**	46	42.49	3.51	12.31	0.29	**NP**	58	57.24	0.76	0.57	0.01
DUF1687	**PP+SPFP**	54	59.64	-5.64	31.80	0.53	**AP**	39	42.76	-3.76	14.12	0.33	**NP**	67	57.60	9.40	88.30	1.53
DUF3844	**PP+SPFP**	58	60.38	-2.38	5.69	0.09	**AP**	48	43.29	4.71	22.16	0.51	**NP**	56	58.32	-2.32	5.40	0.09
DUF3835	**PP+SPFP**	56	61.50	-5.50	30.28	0.49	**AP**	43	44.09	-1.09	1.20	0.03	**NP**	66	59.40	6.60	43.52	0.73
DUF3602	**PP+SPFP**	61	67.47	-6.47	41.82	0.62	**AP**	45	48.37	-3.37	11.36	0.23	**NP**	75	65.16	9.84	96.76	1.48
DUF3779	**PP+SPFP**	61	70.08	-9.08	82.37	1.18	**AP**	50	50.24	-0.24	0.06	0.00	**NP**	77	67.68	9.32	86.80	1.28

PP+SPFP - includes plant pathogenic fungi (PP), symbionts of plant roots and endophyte (SP) and fungi infecting other fungi (FP);
AP – fungi infecting animals, NP - non-pathogenic fungi, O- observed frequencies, E – expected frequencies.

Table C-5: Frequency table for chi-square test 4.
This table is relevant to the text in Chapter 5, section 5.4.3.1 (Statistical evaluation of DUFs association with fungi lifestyle).

DUF Id	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E
DUF2456	**PP**	23	19.58	3.42	11.73	0.60	**AP**	6	15.94	-9.94	98.89	6.20	**NP**	28	21.48	6.52	42.51	1.98
DUF1965	**PP**	29	24.04	4.96	24.60	1.02	**AP**	14	19.58	-5.58	31.15	1.59	**NP**	27	26.38	0.62	0.39	0.01
DUF3716	**PP**	29	24.73	4.27	18.26	0.74	**AP**	19	20.14	-1.14	1.30	0.06	**NP**	24	27.13	-3.13	9.82	0.36
DUF3129	**PP**	44	27.13	16.87	284.58	10.49	**AP**	16	22.10	-6.10	37.19	1.68	**NP**	19	29.77	-10.77	116.01	3.90
DUF2434	**PP**	26	27.13	-1.13	1.28	0.05	**AP**	26	22.10	3.90	15.22	0.69	**NP**	27	29.77	-2.77	7.68	0.26
DUF3176	**PP**	34	27.13	6.87	47.19	1.74	**AP**	18	22.10	-4.10	16.80	0.76	**NP**	27	29.77	-2.77	7.68	0.26
DUF3517	**PP**	35	31.25	3.75	14.05	0.45	**AP**	30	25.46	4.54	20.65	0.81	**NP**	26	34.29	-8.29	68.77	2.01
DUF4045	**PP**	36	32.63	3.37	11.39	0.35	**AP**	30	26.57	3.43	11.74	0.44	**NP**	29	35.80	-6.80	46.25	1.29
DUF4048	**PP**	39	32.97	6.03	36.38	1.10	**AP**	29	26.85	2.15	4.61	0.17	**NP**	28	36.18	-8.18	66.87	1.85
DUF3636	**PP**	38	33.66	4.34	18.87	0.56	**AP**	30	27.41	2.59	6.69	0.24	**NP**	30	36.93	-6.93	48.04	1.30
DUF3807	**PP**	39	34.00	5.00	25.01	0.74	**AP**	27	27.69	-0.69	0.48	0.02	**NP**	33	37.31	-4.31	18.56	0.50
DUF3984	**PP**	39	34.00	5.00	25.01	0.74	**AP**	31	27.69	3.31	10.94	0.39	**NP**	29	37.31	-8.31	69.02	1.85
DUF2014	**PP**	41	35.37	5.63	31.67	0.90	**AP**	27	28.81	-1.81	3.28	0.11	**NP**	35	38.82	-3.82	14.56	0.38
DUF2457	**PP**	41	36.40	4.60	21.13	0.58	**AP**	32	29.65	2.35	5.52	0.19	**NP**	33	39.95	-6.95	48.24	1.21
DUF3292	**PP**	47	36.40	10.60	112.30	3.08	**AP**	23	29.65	-6.65	44.24	1.49	**NP**	36	39.95	-3.95	15.57	0.39
DUF1774	**PP**	34	39.15	-5.15	26.53	0.68	**AP**	38	31.89	6.11	37.34	1.17	**NP**	42	42.96	-0.96	0.92	0.02
DUF3328	**PP**	43	39.15	3.85	14.82	0.38	**AP**	31	31.89	-0.89	0.79	0.02	**NP**	40	42.96	-2.96	8.76	0.20
DUF3433	**PP**	44	38.81	5.19	26.97	0.69	**AP**	31	31.61	-0.61	0.37	0.01	**NP**	38	42.58	-4.58	21.01	0.49
DUF4452	**PP**	36	39.84	-3.84	14.72	0.37	**AP**	35	32.45	2.55	6.51	0.20	**NP**	45	43.71	1.29	1.65	0.04
DUF1770	**PP**	43	39.84	3.16	10.00	0.25	**AP**	32	32.45	-0.45	0.20	0.01	**NP**	41	43.71	-2.71	7.37	0.17
DUF3425	**PP**	48	43.96	4.04	16.33	0.37	**AP**	36	35.81	0.19	0.04	0.00	**NP**	44	48.24	-4.24	17.95	0.37
DUF4484	**PP**	41	45.68	-4.68	21.86	0.48	**AP**	39	37.20	1.80	3.23	0.09	**NP**	53	50.12	2.88	8.29	0.17
DUF2011	**PP**	38	47.39	-9.39	88.22	1.86	**AP**	42	38.60	3.40	11.54	0.30	**NP**	58	52.00	6.00	35.94	0.69
DUF3812	**PP**	40	48.42	-8.42	70.94	1.47	**AP**	40	39.44	0.56	0.31	0.01	**NP**	61	53.14	7.86	61.85	1.16
DUF2406	**PP**	41	48.77	-7.77	60.32	1.24	**AP**	38	39.72	-1.72	2.96	0.07	**NP**	63	53.51	9.49	90.02	1.68
DUF1691	**PP**	44	48.08	-4.08	16.64	0.35	**AP**	41	39.16	1.84	3.38	0.09	**NP**	55	52.76	2.24	5.02	0.10
DUF4448	**PP**	39	49.11	-10.11	102.21	2.08	**AP**	39	40.00	-1.00	1.00	0.03	**NP**	65	53.89	11.11	123.45	2.29
DUF3115	**PP**	42	49.80	-7.80	60.79	1.22	**AP**	44	40.56	3.44	11.83	0.29	**NP**	59	54.64	4.36	18.99	0.35
DUF2417	**PP**	47	51.17	-4.17	17.39	0.34	**AP**	40	41.68	-1.68	2.82	0.07	**NP**	62	56.15	5.85	34.22	0.61
DUF4451	**PP**	48	52.20	-4.20	17.64	0.34	**AP**	46	42.52	3.48	12.12	0.29	**NP**	58	57.28	0.72	0.52	0.01
DUF1687	**PP**	45	51.86	-6.86	47.02	0.91	**AP**	39	42.24	-3.24	10.49	0.25	**NP**	67	56.90	10.10	101.93	1.79
DUF3844	**PP**	51	53.23	-2.23	4.98	0.09	**AP**	48	43.36	4.64	21.55	0.50	**NP**	56	58.41	-2.41	5.81	0.10
DUF3835	**PP**	47	53.57	-6.57	43.22	0.81	**AP**	43	43.64	-0.64	0.41	0.01	**NP**	66	58.79	7.21	52.01	0.88
DUF3602	**PP**	53	59.41	-6.41	41.12	0.69	**AP**	45	48.39	-3.39	11.51	0.24	**NP**	75	65.19	9.81	96.15	1.47
DUF3779	**PP**	54	62.16	-8.16	66.58	1.07	**AP**	50	50.63	-0.63	0.40	0.01	**NP**	77	68.21	8.79	77.28	1.13

PP - plant pathogenic fungi, AP – fungi infecting animals, NP - non-pathogenic fungi, O- observed frequencies, E – expected frequencies.

Table C-6: Frequency table for chi-square test 5.
This table is relevant to the text in Chapter 5, section 5.4.3.1 (Statistical evaluation of DUFs association with fungi lifestyle).

DUF Id	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E
DUF2456	**PP**	23	24.32	-1.32	1.73	0.07	**NP**	28	26.68	1.32	1.73	0.06
DUF1965	**PP**	29	26.70	2.30	5.29	0.20	**NP**	27	29.30	-2.30	5.29	0.18
DUF3716	**PP**	29	25.27	3.73	13.91	0.55	**NP**	24	27.73	-3.73	13.91	0.50
DUF3129	**PP**	44	30.04	13.96	194.93	6.49	**NP**	19	32.96	-13.96	194.93	5.91
DUF2434	**PP**	26	25.27	0.73	0.53	0.02	**NP**	27	27.73	-0.73	0.53	0.02
DUF3176	**PP**	34	29.08	4.92	24.16	0.83	**NP**	27	31.92	-4.92	24.16	0.76
DUF3517	**PP**	35	29.08	5.92	34.99	1.20	**NP**	26	31.92	-5.92	34.99	1.10
DUF4045	**PP**	36	30.99	5.01	25.08	0.81	**NP**	29	34.01	-5.01	25.08	0.74
DUF4048	**PP**	39	31.95	7.05	49.77	1.56	**NP**	28	35.05	-7.05	49.77	1.42
DUF3636	**PP**	38	32.42	5.58	31.11	0.96	**NP**	30	35.58	-5.58	31.11	0.87
DUF3807	**PP**	39	34.33	4.67	21.81	0.64	**NP**	33	37.67	-4.67	21.81	0.58
DUF3984	**PP**	39	32.42	6.58	43.27	1.33	**NP**	29	35.58	-6.58	43.27	1.22
DUF2014	**PP**	41	36.24	4.76	22.69	0.63	**NP**	35	39.76	-4.76	22.69	0.57
DUF2457	**PP**	41	35.28	5.72	32.68	0.93	**NP**	33	38.72	-5.72	32.68	0.84
DUF3292	**PP**	47	39.57	7.43	55.14	1.39	**NP**	36	43.43	-7.43	55.14	1.27
DUF1774	**PP**	34	36.24	-2.24	5.00	0.14	**NP**	42	39.76	2.24	5.00	0.13
DUF3328	**PP**	43	39.57	3.43	11.74	0.30	**NP**	40	43.43	-3.43	11.74	0.27
DUF3433	**PP**	44	39.10	4.90	24.03	0.61	**NP**	38	42.90	-4.90	24.03	0.56
DUF4452	**PP**	36	38.62	-2.62	6.87	0.18	**NP**	45	42.38	2.62	6.87	0.16
DUF1770	**PP**	43	40.05	2.95	8.70	0.22	**NP**	41	43.95	-2.95	8.70	0.20
DUF3425	**PP**	48	43.87	4.13	17.09	0.39	**NP**	44	48.13	-4.13	17.09	0.36
DUF4484	**PP**	41	44.82	-3.82	14.59	0.33	**NP**	53	49.18	3.82	14.59	0.30
DUF2011	**PP**	38	45.77	-7.77	60.41	1.32	**NP**	58	50.23	7.77	60.41	1.20
DUF3812	**PP**	40	48.16	-8.16	66.53	1.38	**NP**	61	52.84	8.16	66.53	1.26
DUF2406	**PP**	41	49.59	-8.59	73.74	1.49	**NP**	63	54.41	8.59	73.74	1.36
DUF1691	**PP**	44	47.20	-3.20	10.26	0.22	**NP**	55	51.80	3.20	10.26	0.20
DUF4448	**PP**	39	49.59	-10.59	112.09	2.26	**NP**	65	54.41	10.59	112.09	2.06
DUF3115	**PP**	42	48.16	-6.16	37.90	0.79	**NP**	59	52.84	6.16	37.90	0.72
DUF2417	**PP**	47	51.97	-4.97	24.71	0.48	**NP**	62	57.03	4.97	24.71	0.43
DUF4451	**PP**	48	50.54	-2.54	6.46	0.13	**NP**	58	55.46	2.54	6.46	0.12
DUF1687	**PP**	45	53.40	-8.40	70.58	1.32	**NP**	67	58.60	8.40	70.58	1.20
DUF3844	**PP**	51	51.02	-0.02	0.00	0.00	**NP**	56	55.98	0.02	0.00	0.00
DUF3835	**PP**	47	53.88	-6.88	47.31	0.88	**NP**	66	59.12	6.88	47.31	0.80
DUF3602	**PP**	53	61.03	-8.03	64.49	1.06	**NP**	75	66.97	8.03	64.49	0.96
DUF3779	**PP**	54	62.46	-8.46	71.58	1.15	**NP**	77	68.54	8.46	71.58	1.04

PP - plant pathogenic fungi, NP - non-pathogenic fungi, O- observed frequencies, E – expected frequencies.

Table C-7: Frequency table for chi-square test 6.
This table is relevant to the text in Chapter 5, section 5.4.3.1 (Statistical evaluation of DUFs association with fungi lifestyle).

DUF Id	Life style	O	E	O-E	(O-E)2	(O-E)2/E	Life style	O	E	O-E	(O-E)2	(O-E)2/E
DUF2456	**AllPath**	33	39.04	-6.04	36.47	0.93	**NP**	28.00	21.96	6.04	36.47	1.66
DUF1965	**AllPath**	45	46.08	-1.08	1.16	0.03	**NP**	27.00	25.92	1.08	1.16	0.04
DUF3716	**AllPath**	52	48.64	3.36	11.30	0.23	**NP**	24.00	27.36	-3.36	11.30	0.41
DUF3129	**AllPath**	64	53.12	10.88	118.41	2.23	**NP**	19.00	29.88	-10.88	118.41	3.96
DUF2434	**AllPath**	57	53.76	3.24	10.51	0.20	**NP**	27.00	30.24	-3.24	10.51	0.35
DUF3176	**AllPath**	57	53.76	3.24	10.51	0.20	**NP**	27.00	30.24	-3.24	10.51	0.35
DUF3517	**AllPath**	70	61.44	8.56	73.30	1.19	**NP**	26.00	34.56	-8.56	73.30	2.12
DUF4045	**AllPath**	71	64.00	7.00	49.03	0.77	**NP**	29.00	36.00	-7.00	49.03	1.36
DUF4048	**AllPath**	73	64.64	8.36	69.92	1.08	**NP**	28.00	36.36	-8.36	69.92	1.92
DUF3636	**AllPath**	72	65.28	6.72	45.18	0.69	**NP**	30.00	36.72	-6.72	45.18	1.23
DUF3807	**AllPath**	70	65.92	4.08	16.66	0.25	**NP**	33.00	37.08	-4.08	16.66	0.45
DUF3984	**AllPath**	75	66.56	8.44	71.27	1.07	**NP**	29.00	37.44	-8.44	71.27	1.90
DUF2014	**AllPath**	73	69.12	3.88	15.07	0.22	**NP**	35.00	38.88	-3.88	15.07	0.39
DUF2457	**AllPath**	77	70.40	6.60	43.59	0.62	**NP**	33.00	39.60	-6.60	43.59	1.10
DUF3292	**AllPath**	75	71.04	3.96	15.70	0.22	**NP**	36.00	39.96	-3.96	15.70	0.39
DUF1774	**AllPath**	77	76.16	0.84	0.71	0.01	**NP**	42.00	42.84	-0.84	0.71	0.02
DUF3328	**AllPath**	79	76.16	2.84	8.08	0.11	**NP**	40.00	42.84	-2.84	8.08	0.19
DUF3433	**AllPath**	81	76.16	4.84	23.45	0.31	**NP**	38.00	42.84	-4.84	23.45	0.55
DUF4452	**AllPath**	75	76.80	-1.80	3.23	0.04	**NP**	45.00	43.20	1.80	3.23	0.07
DUF1770	**AllPath**	80	77.44	2.56	6.57	0.08	**NP**	41.00	43.56	-2.56	6.57	0.15
DUF3425	**AllPath**	89	85.12	3.88	15.07	0.18	**NP**	44.00	47.88	-3.88	15.07	0.31
DUF4484	**AllPath**	86	88.96	-2.96	8.75	0.10	**NP**	53.00	50.04	2.96	8.75	0.17
DUF2011	**AllPath**	85	91.52	-6.52	42.48	0.46	**NP**	58.00	51.48	6.52	42.48	0.83
DUF3812	**AllPath**	86	94.08	-8.08	65.24	0.69	**NP**	61.00	52.92	8.08	65.24	1.23
DUF2406	**AllPath**	85	94.72	-9.72	94.43	1.00	**NP**	63.00	53.28	9.72	94.43	1.77
DUF1691	**AllPath**	93	94.72	-1.72	2.95	0.03	**NP**	55.00	53.28	1.72	2.95	0.06
DUF4448	**AllPath**	85	96.00	-11.00	120.94	1.26	**NP**	65.00	54.00	11.00	120.94	2.24
DUF3115	**AllPath**	92	96.64	-4.64	21.50	0.22	**NP**	59.00	54.36	4.64	21.50	0.40
DUF2417	**AllPath**	93	99.20	-6.20	38.40	0.39	**NP**	62.00	55.80	6.20	38.40	0.69
DUF4451	**AllPath**	101	101.76	-0.76	0.57	0.01	**NP**	58.00	57.24	0.76	0.57	0.01
DUF1687	**AllPath**	93	102.40	-9.40	88.30	0.86	**NP**	67.00	57.60	9.40	88.30	1.53
DUF3844	**AllPath**	106	103.68	2.32	5.40	0.05	**NP**	56.00	58.32	-2.32	5.40	0.09
DUF3835	**AllPath**	99	105.60	-6.60	43.52	0.41	**NP**	66.00	59.40	6.60	43.52	0.73
DUF3602	**AllPath**	106	115.84	-9.84	96.76	0.84	**NP**	75.00	65.16	9.84	96.76	1.48
DUF3779	**AllPath**	111	120.32	-9.32	86.80	0.72	**NP**	77.00	67.68	9.32	86.80	1.28

AllPath– includes fungi lifestyle groups: plant pathogenic fungi (PP), symbionts of plant roots and endophyte (SP), fungi infecting other fungi (FP) and fungi infecting animals (AP); NP - non-pathogenic fungi, O- observed frequencies; E – expected frequencies.

Table C-8 Domains and the DUFs content within all the connected components of the domain-association network.
This table is relevant to the text in Chapter 5, section 5.4.4 (*Fusarium graminearum* domain-association network).

1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
CC ID	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**
1	705	40.35	40	5.67	39.60	**23**	5	0.29	0	0.00	0.00	**45**	4	0.23	1	25.00	0.99
2	15	0.86	0	0.00	0.00	**24**	5	0.29	0	0.00	0.00	**46**	4	0.23	0	0.00	0.00
3	14	0.80	1	7.14	0.99	**25**	5	0.29	0	0.00	0.00	**47**	4	0.23	0	0.00	0.00
4	12	0.69	0	0.00	0.00	**26**	5	0.29	0	0.00	0.00	**48**	4	0.23	0	0.00	0.00
5	11	0.63	0	0.00	0.00	**27**	5	0.29	0	0.00	0.00	**49**	4	0.23	0	0.00	0.00
6	11	0.63	0	0.00	0.00	**28**	5	0.29	0	0.00	0.00	**50**	4	0.23	0	0.00	0.00
7	11	0.63	0	0.00	0.00	**29**	5	0.29	0	0.00	0.00	**51**	4	0.23	0	0.00	0.00
8	9	0.52	0	0.00	0.00	**30**	4	0.23	0	0.00	0.00	**52**	4	0.23	0	0.00	0.00
9	8	0.46	0	0.00	0.00	**31**	4	0.23	1	25.00	0.99	**53**	4	0.23	0	0.00	0.00
10	7	0.40	0	0.00	0.00	**32**	4	0.23	1	25.00	0.99	**54**	4	0.23	0	0.00	0.00
11	7	0.40	1	14.29	0.99	**33**	4	0.23	1	25.00	0.99	**55**	4	0.23	0	0.00	0.00
12	7	0.40	0	0.00	0.00	**34**	4	0.23	0	0.00	0.00	**56**	4	0.23	1	25.00	0.99
13	7	0.40	0	0.00	0.00	**35**	4	0.23	0	0.00	0.00	**57**	4	0.23	1	25.00	0.99
14	7	0.40	0	0.00	0.00	**36**	4	0.23	1	25.00	0.99	**58**	3	0.17	0	0.00	0.00
15	7	0.40	1	14.29	0.99	**37**	4	0.23	1	25.00	0.99	**59**	3	0.17	0	0.00	0.00
16	6	0.34	0	0.00	0.00	**38**	4	0.23	0	0.00	0.00	**60**	3	0.17	0	0.00	0.00
17	6	0.34	0	0.00	0.00	**39**	4	0.23	0	0.00	0.00	**61**	3	0.17	0	0.00	0.00
18	6	0.34	1	16.67	0.99	**40**	4	0.23	0	0.00	0.00	**62**	3	0.17	0	0.00	0.00
19	6	0.34	1	16.67	0.99	**41**	4	0.23	0	0.00	0.00	**63**	3	0.17	0	0.00	0.00
20	6	0.34	1	16.67	0.99	**42**	4	0.23	0	0.00	0.00	**64**	3	0.17	0	0.00	0.00
21	6	0.34	0	0.00	0.00	**43**	4	0.23	3	75.00	2.97	**65**	3	0.17	0	0.00	0.00
22	6	0.34	0	0.00	0.00	**44**	4	0.23	0	0.00	0.00	**66**	3	0.17	0	0.00	0.00

Table C-8 continues.

1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
CC ID	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**
67	3	0.17	1	33.33	0.99	**89**	3	0.17	0	0.00	0.00	**111**	3	0.17	0	0.00	0.00
68	3	0.17	0	0.00	0.00	**90**	3	0.17	0	0.00	0.00	**112**	3	0.17	0	0.00	0.00
69	3	0.17	0	0.00	0.00	**91**	3	0.17	0	0.00	0.00	**113**	3	0.17	1	33.33	0.99
70	3	0.17	0	0.00	0.00	**92**	3	0.17	0	0.00	0.00	**114**	3	0.17	0	0.00	0.00
71	3	0.17	0	0.00	0.00	**93**	3	0.17	0	0.00	0.00	**115**	3	0.17	1	33.33	0.99
72	3	0.17	0	0.00	0.00	**94**	3	0.17	0	0.00	0.00	**116**	3	0.17	0	0.00	0.00
73	3	0.17	0	0.00	0.00	**95**	3	0.17	0	0.00	0.00	**117**	3	0.17	0	0.00	0.00
74	3	0.17	0	0.00	0.00	**96**	3	0.17	0	0.00	0.00	**118**	3	0.17	0	0.00	0.00
75	3	0.17	0	0.00	0.00	**97**	3	0.17	0	0.00	0.00	**119**	3	0.17	0	0.00	0.00
76	3	0.17	0	0.00	0.00	**98**	3	0.17	0	0.00	0.00	**120**	2	0.11	0	0.00	0.00
77	3	0.17	0	0.00	0.00	**99**	3	0.17	0	0.00	0.00	**121**	2	0.11	0	0.00	0.00
78	3	0.17	0	0.00	0.00	**100**	3	0.17	0	0.00	0.00	**122**	2	0.11	0	0.00	0.00
79	3	0.17	1	33.33	0.99	**101**	3	0.17	0	0.00	0.00	**123**	2	0.11	0	0.00	0.00
80	3	0.17	0	0.00	0.00	**102**	3	0.17	0	0.00	0.00	**124**	2	0.11	0	0.00	0.00
81	3	0.17	2	66.67	1.98	**103**	3	0.17	0	0.00	0.00	**125**	2	0.11	0	0.00	0.00
82	3	0.17	1	33.33	0.99	**104**	3	0.17	0	0.00	0.00	**126**	2	0.11	0	0.00	0.00
83	3	0.17	0	0.00	0.00	**105**	3	0.17	0	0.00	0.00	**127**	2	0.11	0	0.00	0.00
84	3	0.17	0	0.00	0.00	**106**	3	0.17	0	0.00	0.00	**128**	2	0.11	0	0.00	0.00
85	3	0.17	0	0.00	0.00	**107**	3	0.17	0	0.00	0.00	**129**	2	0.11	0	0.00	0.00
86	3	0.17	0	0.00	0.00	**108**	3	0.17	0	0.00	0.00	**130**	2	0.11	0	0.00	0.00
87	3	0.17	0	0.00	0.00	**109**	3	0.17	0	0.00	0.00	**131**	2	0.11	0	0.00	0.00
88	3	0.17	0	0.00	0.00	**110**	3	0.17	0	0.00	0.00	**132**	2	0.11	1	50.00	0.99

Table C-8 continues.

1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
CC ID	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**
133	2	0.11	0	0.00	0.00	**155**	2	0.11	0	0.00	0.00	**177**	2	0.11	0	0.00	0.00
134	2	0.11	0	0.00	0.00	**156**	2	0.11	0	0.00	0.00	**178**	2	0.11	0	0.00	0.00
135	2	0.11	0	0.00	0.00	**157**	2	0.11	0	0.00	0.00	**179**	2	0.11	0	0.00	0.00
136	2	0.11	0	0.00	0.00	**158**	2	0.11	0	0.00	0.00	**180**	2	0.11	2	100.00	1.98
137	2	0.11	0	0.00	0.00	**159**	2	0.11	0	0.00	0.00	**181**	2	0.11	0	0.00	0.00
138	2	0.11	0	0.00	0.00	**160**	2	0.11	0	0.00	0.00	**182**	2	0.11	0	0.00	0.00
139	2	0.11	0	0.00	0.00	**161**	2	0.11	0	0.00	0.00	**183**	2	0.11	0	0.00	0.00
140	2	0.11	0	0.00	0.00	**162**	2	0.11	0	0.00	0.00	**184**	2	0.11	0	0.00	0.00
141	2	0.11	0	0.00	0.00	**163**	2	0.11	0	0.00	0.00	**185**	2	0.11	0	0.00	0.00
142	2	0.11	0	0.00	0.00	**164**	2	0.11	1	50.00	0.99	**186**	2	0.11	1	50.00	0.99
143	2	0.11	0	0.00	0.00	**165**	2	0.11	0	0.00	0.00	**187**	2	0.11	0	0.00	0.00
144	2	0.11	0	0.00	0.00	**166**	2	0.11	0	0.00	0.00	**188**	2	0.11	0	0.00	0.00
145	2	0.11	0	0.00	0.00	**167**	2	0.11	0	0.00	0.00	**189**	2	0.11	0	0.00	0.00
146	2	0.11	0	0.00	0.00	**168**	2	0.11	0	0.00	0.00	**190**	2	0.11	0	0.00	0.00
147	2	0.11	0	0.00	0.00	**169**	2	0.11	0	0.00	0.00	**191**	2	0.11	0	0.00	0.00
148	2	0.11	0	0.00	0.00	**170**	2	0.11	0	0.00	0.00	**192**	2	0.11	0	0.00	0.00
149	2	0.11	0	0.00	0.00	**171**	2	0.11	0	0.00	0.00	**193**	2	0.11	0	0.00	0.00
150	2	0.11	0	0.00	0.00	**172**	2	0.11	0	0.00	0.00	**194**	2	0.11	0	0.00	0.00
151	2	0.11	0	0.00	0.00	**173**	2	0.11	0	0.00	0.00	**195**	2	0.11	2	100.00	1.98
152	2	0.11	0	0.00	0.00	**174**	2	0.11	0	0.00	0.00	**196**	2	0.11	0	0.00	0.00
153	2	0.11	0	0.00	0.00	**175**	2	0.11	0	0.00	0.00	**197**	2	0.11	0	0.00	0.00
154	2	0.11	0	0.00	0.00	**176**	2	0.11	2	100.00	1.98	**198**	2	0.11	0	0.00	0.00

Table C-8 continues.

1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
CC ID	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**
199	2	0.11	0	0.00	0.00	**221**	2	0.11	0	0.00	0.00	**243**	2	0.11	0	0.00	0.00
200	2	0.11	0	0.00	0.00	**222**	2	0.11	0	0.00	0.00	**244**	2	0.11	0	0.00	0.00
201	2	0.11	0	0.00	0.00	**223**	2	0.11	0	0.00	0.00	**245**	2	0.11	0	0.00	0.00
202	2	0.11	0	0.00	0.00	**224**	2	0.11	0	0.00	0.00	**246**	2	0.11	0	0.00	0.00
203	2	0.11	0	0.00	0.00	**225**	2	0.11	0	0.00	0.00	**247**	2	0.11	0	0.00	0.00
204	2	0.11	0	0.00	0.00	**226**	2	0.11	0	0.00	0.00	**248**	2	0.11	0	0.00	0.00
205	2	0.11	1	50.00	0.99	**227**	2	0.11	0	0.00	0.00	**249**	2	0.11	1	50.00	0.99
206	2	0.11	0	0.00	0.00	**228**	2	0.11	0	0.00	0.00	**250**	2	0.11	2	100.00	1.98
207	2	0.11	0	0.00	0.00	**229**	2	0.11	0	0.00	0.00	**251**	2	0.11	0	0.00	0.00
208	2	0.11	0	0.00	0.00	**230**	2	0.11	0	0.00	0.00	**252**	2	0.11	0	0.00	0.00
209	2	0.11	0	0.00	0.00	**231**	2	0.11	2	100.00	1.98	**253**	2	0.11	0	0.00	0.00
210	2	0.11	0	0.00	0.00	**232**	2	0.11	0	0.00	0.00	**254**	2	0.11	0	0.00	0.00
211	2	0.11	0	0.00	0.00	**233**	2	0.11	0	0.00	0.00	**255**	2	0.11	0	0.00	0.00
212	2	0.11	0	0.00	0.00	**234**	2	0.11	1	50.00	0.99	**256**	2	0.11	0	0.00	0.00
213	2	0.11	0	0.00	0.00	**235**	2	0.11	0	0.00	0.00	**257**	2	0.11	0	0.00	0.00
214	2	0.11	0	0.00	0.00	**236**	2	0.11	0	0.00	0.00	**258**	2	0.11	0	0.00	0.00
215	2	0.11	0	0.00	0.00	**237**	2	0.11	0	0.00	0.00	**259**	2	0.11	1	50.00	0.99
216	2	0.11	0	0.00	0.00	**238**	2	0.11	0	0.00	0.00	**260**	2	0.11	0	0.00	0.00
217	2	0.11	0	0.00	0.00	**239**	2	0.11	0	0.00	0.00	**261**	2	0.11	0	0.00	0.00
218	2	0.11	0	0.00	0.00	**240**	2	0.11	0	0.00	0.00	**262**	2	0.11	0	0.00	0.00
219	2	0.11	0	0.00	0.00	**241**	2	0.11	0	0.00	0.00	**263**	2	0.11	0	0.00	0.00
220	2	0.11	0	0.00	0.00	**242**	2	0.11	0	0.00	0.00	**264**	2	0.11	0	0.00	0.00

Table C-8 continues.

1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
CC ID	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**
265	2	0.11	0	0.00	0.00	**287**	2	0.11	0	0.00	0.00	**309**	2	0.11	0	0.00	0.00
266	2	0.11	0	0.00	0.00	**288**	2	0.11	1	50.00	0.99	**310**	2	0.11	0	0.00	0.00
267	2	0.11	0	0.00	0.00	**289**	2	0.11	0	0.00	0.00	**311**	2	0.11	0	0.00	0.00
268	2	0.11	0	0.00	0.00	**290**	2	0.11	0	0.00	0.00	**312**	2	0.11	0	0.00	0.00
269	2	0.11	0	0.00	0.00	**291**	2	0.11	1	50.00	0.99	**313**	2	0.11	2	100.00	1.98
270	2	0.11	0	0.00	0.00	**292**	2	0.11	2	100.00	1.98	**314**	2	0.11	2	100.00	1.98
271	2	0.11	0	0.00	0.00	**293**	2	0.11	0	0.00	0.00	**315**	2	0.11	0	0.00	0.00
272	2	0.11	0	0.00	0.00	**294**	2	0.11	0	0.00	0.00	**316**	2	0.11	0	0.00	0.00
273	2	0.11	0	0.00	0.00	**295**	2	0.11	0	0.00	0.00	**317**	2	0.11	0	0.00	0.00
274	2	0.11	0	0.00	0.00	**296**	2	0.11	0	0.00	0.00	**318**	2	0.11	0	0.00	0.00
275	2	0.11	0	0.00	0.00	**297**	2	0.11	0	0.00	0.00	**319**	2	0.11	0	0.00	0.00
276	2	0.11	0	0.00	0.00	**298**	2	0.11	0	0.00	0.00	**320**	2	0.11	0	0.00	0.00
277	2	0.11	0	0.00	0.00	**299**	2	0.11	0	0.00	0.00	**321**	2	0.11	0	0.00	0.00
278	2	0.11	0	0.00	0.00	**300**	2	0.11	1	50.00	0.99	**322**	2	0.11	0	0.00	0.00
279	2	0.11	0	0.00	0.00	**301**	2	0.11	0	0.00	0.00	**323**	2	0.11	0	0.00	0.00
280	2	0.11	0	0.00	0.00	**302**	2	0.11	0	0.00	0.00	**324**	2	0.11	0	0.00	0.00
281	2	0.11	0	0.00	0.00	**303**	2	0.11	0	0.00	0.00	**325**	2	0.11	0	0.00	0.00
282	2	0.11	1	50.00	0.99	**304**	2	0.11	0	0.00	0.00	**326**	2	0.11	0	0.00	0.00
283	2	0.11	0	0.00	0.00	**305**	2	0.11	1	50.00	0.99	**327**	2	0.11	0	0.00	0.00
284	2	0.11	0	0.00	0.00	**306**	2	0.11	0	0.00	0.00	**328**	2	0.11	1	50.00	0.99
285	2	0.11	0	0.00	0.00	**307**	2	0.11	0	0.00	0.00	**329**	2	0.11	1	50.00	0.99
286	2	0.11	0	0.00	0.00	**308**	2	0.11	0	0.00	0.00	**330**	2	0.11	0	0.00	0.00

Table C-8 continues.

1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
CC ID	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**	**CC ID**	**Domains No**	**Domains [%]**	**DUFs No**	**DUFs in CC [%]**	**DUFs % of total DUFs**
331	2	0.11	0	0.00	0.00	**350**	2	0.11	0	0.00	0.00	**369**	2	0.11	0	0.00	0.00
332	2	0.11	0	0.00	0.00	**351**	2	0.11	0	0.00	0.00	**370**	2	0.11	0	0.00	0.00
333	2	0.11	0	0.00	0.00	**352**	2	0.11	0	0.00	0.00	**371**	2	0.11	0	0.00	0.00
334	2	0.11	0	0.00	0.00	**353**	2	0.11	0	0.00	0.00	**372**	2	0.11	0	0.00	0.00
335	2	0.11	0	0.00	0.00	**354**	2	0.11	0	0.00	0.00	**373**	2	0.11	0	0.00	0.00
336	2	0.11	0	0.00	0.00	**355**	2	0.11	0	0.00	0.00	**374**	2	0.11	0	0.00	0.00
337	2	0.11	0	0.00	0.00	**356**	2	0.11	0	0.00	0.00	**375**	2	0.11	0	0.00	0.00
338	2	0.11	1	50.00	0.99	**357**	2	0.11	0	0.00	0.00	**376**	2	0.11	0	0.00	0.00
339	2	0.11	0	0.00	0.00	**358**	2	0.11	0	0.00	0.00	**377**	2	0.11	0	0.00	0.00
340	2	0.11	0	0.00	0.00	**359**	2	0.11	0	0.00	0.00	**378**	2	0.11	0	0.00	0.00
341	2	0.11	0	0.00	0.00	**360**	2	0.11	0	0.00	0.00	**379**	2	0.11	0	0.00	0.00
342	2	0.11	0	0.00	0.00	**361**	2	0.11	0	0.00	0.00	**380**	2	0.11	0	0.00	0.00
343	2	0.11	0	0.00	0.00	**362**	2	0.11	0	0.00	0.00	**381**	2	0.11	0	0.00	0.00
344	2	0.11	0	0.00	0.00	**363**	2	0.11	0	0.00	0.00	**382**	2	0.11	0	0.00	0.00
345	2	0.11	0	0.00	0.00	**364**	2	0.11	0	0.00	0.00	**383**	2	0.11	0	0.00	0.00
346	2	0.11	0	0.00	0.00	**365**	2	0.11	0	0.00	0.00	**384**	2	0.11	2	100.00	1.98
347	2	0.11	2	100.00	1.98	**366**	2	0.11	0	0.00	0.00	**385**	2	0.11	0	0.00	0.00
348	2	0.11	2	100.00	1.98	**367**	2	0.11	0	0.00	0.00	**386**	2	0.11	0	0.00	0.00
349	2	0.11	0	0.00	0.00	**368**	2	0.11	0	0.00	0.00						

1 CC ID – connected component number (id), 2 Domains No – number of pfam domains including DUFs in the connected component, 3 Domains [%] – percentage of the number of domains in the given connected component reflecting the total number of the domain in the network (where the total number of domains in the network equals to 1747), 4 DUFs No – number of Domains of Unknown Function in the given connected component, 5 DUFs in CC [%] – percentage of pfam domains that are Domains of Unknown Function in the given connected component, 6 DUFs % of total DUFs – percentage of the number of Domains of Unknown Function in the given connected component reflecting the total number of Domains of Unknown Function in the network (where the total number of Domains of Unknown Function in the network is equals to 101).

Appendix D.

Resolving the pfam domains overlapping problem in *F. graminearum*

The information in Appendix D is relevant to Chapter 5, section 5.3.1 (Identification of domain composition in *Fusarium graminearum* genome).

I. Introduction

In order to resolve the pfam domains overlapping problem in FG proteins, a python script was written where rules to solve this issue were adopted from previous work (Seidl et al., 2011). However, using the python script, it was not possible to resolve the issue of overlapping in 12 FG proteins. This is because of the high level of overlap between multiple domains within the protein. Therefore, a manual approach was used.

Initially, conserved domains search for each of those 12 FG proteins was performed with the aid of Conserved Domain Database (CDD) version 3.10 (on 4-5 June 2013). For that purpose each of 12 FG protein sequences in FASTA format from MIPS FGDB were submitted separately to CD-search of CDD ((www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) using defaults parameters. The result of each search was displayed in both output modes: concise and full displays. **Concise display** shows only the best scoring domain model, in each hit category listed below except non-specific hits, for each region on the query sequence, whereas **full display** shows all domain models, in each hit category below, that meet or exceed the RPS-BLAST threshold for statistical significance.

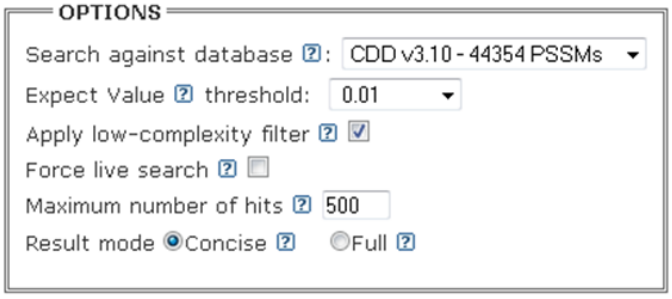


Figure D-1 Defaults parameter used in CD-search (here for obtaining results in concise display)

In majority of CD-search run (7 out of 12 proteins) I did not obtain any hit for pfam domain in concise mode output. Moreover, hits in concise mode for 5 FG proteins contradicted with the score of their domains (generated via Python script) and the best e-value for their domains (generated via Hmmer3 run).

This outcome led to the change of the steps taken in finding the best domain from those overlapping within the given protein (see below).

II. Methodology used

Firstly, the score for each domain is taken into account. If the score is equal for all or some of the domains, the pairwise score comparison for those is performed. If the pairwise score comparison is not able to resolve the overlapping (is not able to point out the best domain) the domain with the best e-value (from Hmmer3 run) is chosen from the group of domains with the same score and the same results of pairwise score comparison. The example of scoring table together with the information generated with the aid of hmmer3 run is presented in Tables D-1 A and B.

1. Steps in finding the best solution for overlapping domains in the example from the Table D-1.

a. All of domains overlap with each other (Table D-1 A, column c and d). The best equal score (1 x 0, 4 x 1) was calculated for three domains: PF08241, PF13847 and PF13489 (Table D-1 B). The next step is to perform pairwise comparison of these three domains.

b. Pairwise score comparison of 3 domains from point 1:

- Comparison of domain 1 and domain 3 results in score equals to 0 (row 1, column 3) – exclusion of domain 1
- Comparison of domain 3 and domain 1 results in score equals to 1 (row 3, column 1) – domain 3 has a chance to be the best one
- Comparison of domain 6 and domain 3 results in score equals to 1 (row 6, column 3) – exclusion of domain 3 and chance for domain 6 to be the best one.
- Comparison of domain 6 and domain 1 results in score equals to 0 (row 6, column 1)

Thus, overlapping is not resolved via pairwise score comparison. Then, the domain (from these three domains with the same score) with the best e-value is chosen. In the example above domain PF08241 was chosen as the best one.

c. The last step: the domain with the lowest e-value (hmmer run) is chosen. In the above example the domain PF08241 has the lowest e-value. (Table D-1 A, columns g and h). Therefore, domain PF08241 is chosen.

Table D-1 Example of finding the best domain based first on the score, then the pairwise score and finally on the best e-value.

A. The information from hmmer run

a	b	c	d	e	f	g	h
Pfam ID	Domain description	Start	End	End -Start	Score	e-value	-Log10 (e-value)
PF08241	Methyltransferase domain	40	146	106	55.48	3.00E-13	12.5229
PF08242	Methyltransferase domain	40	144	104	49.24	2.20E-11	10.6576
PF13847	Methyltransferase domain	33	196	163	40.11	1.20E-08	7.9208
PF13649	Methyltransferase domain	39	142	103	34.17	7.70E-07	6.1135
PF12847	Methyltransferase domain	35	149	114	34.07	8.20E-07	6.0862
PF13489	Methyltransferase domain	8	202	194	32.75	0.000002	5.699

B. Scores: (-1) – domain does not overlap, (1) – domain overlaps and it is the best choice (win), (0) – domain overlaps and is not taken into account.

No	Pfam ID	PF08241	PF08242	PF13847	PF13649	PF12847	PF13489
1	PF08241	-1	1	0	1	1	1
2	PF08242	0	-1	0	1	0	0
3	PF13847	1	1	-1	1	1	0
4	PF13649	0	0	0	-1	0	0
5	PF12847	0	1	0	1	-1	0
6	PF13489	0	1	1	1	1	-1

III. Analysis of pfam domains within 12 FG proteins, manually solving overlapping domains and justification

1. FGSG_16446 related to transcription factor Pig1p

Table D-2 Overlapping domains in FGSG_16446

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF00096	Zinc finger, C2H2 type	10	32	22	61.96	3.30E-15	14.4815
PF00096	Zinc finger, C2H2 type	38	60	22	61.96	3.30E-15	14.4815
PF13894	C2H2-type zinc finger	10	33	23	49.45	1.90E-11	10.7212
PF13894	C2H2-type zinc finger	38	61	23	49.45	1.90E-11	10.7212
PF13465	Zinc-finger double domain	24	49	25	34.52	6.00E-07	6.2218
PF00172	Fungal Zn(2)-Cys(6) binuclear cluster domain	82	119	37	28.41	0.000042	4.3768
PF04082	Fungal-specific transcription factor domain	360	621	261	26.01	0.000219	3.6596

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost.

	PF00096	PF00096	PF13894	PF13894	PF13465	PF00172	PF04082
PF00096	-1	-1	0	-1	1	-1	-1
PF00096	-1	-1	-1	0	1	-1	-1
PF13894	1	-1	-1	-1	0	-1	-1
PF13894	-1	1	-1	-1	0	-1	-1
PF13465	0	0	1	1	-1	-1	-1
PF00172	-1	-1	-1	-1	-1	-1	-1
PF04082	-1	-1	-1	-1	-1	-1	-1

a. Proposed domains in FGSG_16446:

PF13894 (10-33), **PF13894** (38-61), **PF00172** (82-119), **PF04082** (360-621)

b. Justification:

Based on Tables D-2 A and B, two sets of domains overlap (Set one: PF00096 (10-32), PF13894 (10-33), PF13465 (24-49); and set two: PF00096 (38-60), PF13894 (38-61), PF13465 (24-49)) and two domains are not overlapping: PF00172 (82-119), PF04082 (360-621).

Solving the domains overlapping in set one:

Based on the scoring table (Table D-2 B), domain PF13465 is not considered as it has lower scoring than two other domains. As domains PF00096 and, PF13894 have the same score (Table D-2 B), pairwise comparison of these two domains was performed. Comparing PF00096 (10-32) domain with PF13894 (10-33) domain the score equals to 0, while comparing PF13894 (10-33) domain with PF00096 (10-32) domain the obtained score is equal to 1. Thus, a comparison of domain PF00096 (10-32) with domain PF13894 (10-33) ends in favour of domain PF13894 (10-33) (based on the pairwise score comparison).

Solving the domains overlapping in set two:

Based on the scoring table (Table D-2 B), domain PF13465 is not considered as it has lower scoring than two other domains in the set. As domains PF00096 (38-60) and PF13894 (38-61) have the same score (Table D-2 B) the pairwise comparison of these two domains was performed. Comparing PF00096 (38-60) domain with PF13894 (38-61) domain the score equals to 0, while comparing PF13894 (38-61) domain with PF00096 (38-60) domain, the obtained score is equal to 1. Thus, comparison of domain PF00096 (38-60) with domain PF13894 (38-61) ends in favour of domain PF13894 (38-61) (based on the pairwise score comparison).

2. FGSG_15775 Hypothetical protein

Table D-3 Overlapping domains in FGSG_15775.

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF08241	Methyltransferase domain	40	146	106	55.48	3.00E-13	12.5229
PF08242	Methyltransferase domain	40	144	104	49.24	2.20E-11	10.6576
PF13847	Methyltransferase domain	33	196	163	40.11	1.20E-08	7.9208
PF13649	Methyltransferase domain	39	142	103	34.17	7.70E-07	6.1135
PF12847	Methyltransferase domain	35	149	114	34.07	8.20E-07	6.0862
PF13489	Methyltransferase domain	8	202	194	32.75	0.000002	5.699

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost.

	PF08241	PF08242	PF13847	PF13649	PF12847	PF13489
PF08241	-1	1	0	1	1	1
PF08242	0	-1	0	1	0	0
PF13847	1	1	-1	1	1	0
PF13649	0	0	0	-1	0	0
PF12847	0	1	0	1	-1	0
PF13489	0	1	1	1	1	-1

a. Proposed domains in FGSG_15775

PF08241 (40-146)

b. Justification

Based on the information in Tables D-3 A and B, every domain overlaps with each other. However, domains PF08241 (40-146), PF13847 (33-196), and PF13489 (8-202) have the equal and the highest scoring reported in Table D-3 B. Although pairwise comparison was performed on these three domains, the overlapping was not resolved as the same scores were obtained for all possible pairs of these three domains. Finally, based on the best e-value, domain PF08241 (40-146) was chosen.

3. FGSG_16504 hypothetical protein

Table D-4 Overlapping domains in FGSG_16504

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF02190	ATP-dependent protease La (LON) domain	340	583	243	99.03	2.30E-26	25.6383
PF13923	Zinc finger, C3HC4 type (RING finger)	72	124	52	66.15	1.80E-16	15.7447
PF13923	Zinc finger, C3HC4 type (RING finger)	249	286	37	66.15	1.80E-16	15.7447
PF13639	Ring finger domain	247	287	40	56.46	1.50E-13	12.8239
PF14634	zinc-RING finger domain	248	288	40	55.95	2.10E-13	12.6778
PF00097	Zinc finger, C3HC4 type (RING finger)	249	286	37	55.08	3.90E-13	12.4089
PF15227	zinc finger of C3HC4-type, RING	249	286	37	53.19	1.40E-12	11.8539
PF13920	Zinc finger, C3HC4 type (RING finger)	245	293	48	48.07	5.00E-11	10.301
PF13445	RING-type zinc-finger	249	284	35	43	1.70E-09	8.7696

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost.

	PF02190	PF13923	PF13923	PF13639	PF14634	PF00097	PF15227	PF13920	PF13445
PF02190	-1	-1	-1	-1	-1	-1	-1	-1	-1
PF13923	-1	-1	-1	-1	-1	-1	-1	-1	-1
PF13923	-1	-1	-1	0	0	1	1	1	1
PF13639	-1	-1	1	-1	1	1	1	0	1
PF14634	-1	-1	1	0	-1	1	1	0	1
PF00097	-1	-1	0	0	0	-1	1	0	1
PF15227	-1	-1	0	0	0	0	-1	0	1
PF13920	-1	-1	0	1	1	1	1	-1	1
PF13445	-1	-1	0	0	0	0	0	0	-1

a. Proposed domains in FGSG_16504

PF13923 (72-124), **PF13920** (245-293), **PF02190** (340-583)

b. Justification

Domains PF02190 (340-583) and PF13923 (72-124) do not overlap with any of the domains within the protein sequence, whereas domains PF13923 (249-286), PF13639 (247-287), PF14634 (248-288), PF00097 (249-286), PF15227 (249-286), PF13920 (245-293) and PF13445 (249-284) overlap with each other. Domains PF13639 (247-287) and PF13920 (245-293) have the equal highest score. After pairwise comparison of both domains PF13920 (245-293) was chosen as the best one.

4. FGSG_06199 hypothetical protein

Table D-5 Overlapping domains in FGSG_06199

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF08242	Methyltransferase domain	156	248	92	52.4	2.50E-12	11.6021
PF08241	Methyltransferase domain	156	250	94	52.3	2.70E-12	11.5686
PF13847	Methyltransferase domain	149	317	168	44	8.40E-10	9.0757
PF13649	Methyltransferase domain	155	246	91	39.55	1.80E-08	7.7447
PF12847	Methyltransferase domain	151	253	102	39.47	1.90E-08	7.7212
PF13489	Methyltransferase domain	128	349	221	35.7	2.70E-07	6.5686

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost.

	PF08242	PF08241	PF13847	PF13649	PF12847	PF13489
PF08242	-1	0	0	1	0	1
PF08241	1	-1	0	1	0	0
PF13847	1	1	-1	1	1	0
PF13649	0	0	0	-1	0	0
PF12847	1	1	0	1	-1	0
PF13489	0	1	1	1	1	-1

a. Proposed domains in FGSG_06199

PF13847 (149-317)

b. Justification

All domains overlap with each other. However, both PF13847 (149-317) domain and PF13489 (128-349) domain have the highest scoring (Table D-5 B). Pairwise comparison of these two domains gives the same results. Thus, domain PF13847 (149-317) was chosen as it has the smallest e-value.

5. FGSG_16806 hypothetical protein

Table D-6 Overlapping domains in FGSG_16806

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF00097	Zinc finger, C3HC4 type (RING finger)	16	71	55	47.6	7.00E-11	10.1549
PF13639	Ring finger domain	14	72	58	38.37	4.20E-08	7.3768
PF13923	Zinc finger, C3HC4 type (RING finger)	16	71	55	35.48	3.10E-07	6.5086
PF14634	zinc-RING finger domain	15	73	58	34.52	6.00E-07	6.2218
PF14604	Variant SH3 domain	879	929	50	34.44	6.40E-07	6.1938
PF13445	RING-type zinc-finger	16	69	53	28.62	0.000036	4.4437
PF13920	Zinc finger, C3HC4 type (RING finger)	12	78	66	27.76	0.000065	4.1871
PF00018	SH3 domain	878	925	47	25.41	0.000333	3.4776

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost.

	PF00097	PF13639	PF13923	PF14634	PF14604	PF13445	PF13920	PF00018
PF00097	-1	0	1	0	-1	1	1	-1
PF13639	1	-1	1	1	-1	1	0	-1
PF13923	0	0	-1	0	-1	1	0	-1
PF14634	1	0	1	-1	-1	1	0	-1
PF14604	-1	-1	-1	-1	-1	-1	-1	1
PF13445	0	0	0	0	-1	-1	0	-1
PF13920	0	1	1	1	-1	1	-1	-1
PF00018	-1	-1	-1	-1	0	-1	-1	-1

a. Proposed domains in FGSG_16806

PF13920 (12-78), **PF14604** (879-929)

b. Justification

There are two sets of overlapping domains:

- Set one: PF00097 (16-71), PF13639 (14-72), PF13923 (16-71), PF14634 (15-73), PF13445 (16-69) and PF13920 (12-78)
- Set two: PF14604 (879-929) and PF00018 (878-925).

Solving overlapping in set one:

Domains PF13639 (14-72) and PF13920 (12-78) have the highest score within set one (Table D-6 B). After pairwise scoring comparison of these two domains, domain PF13920 (12-78) was chosen.

Solving overlapping in set two:

Domain PF14604 (879-929) has a higher score than domain PF00018 (878-925). Thus, domain PF14604 (879-929) was chosen.

6. FGSG_16768 related to krueppel protein

Table D- 7 Overlapping domains in FGSG_16768

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF00096	Zinc finger, C2H2 type	149	171	22	57.1	9.40E-14	13.0269
PF00096	Zinc finger, C2H2 type	177	201	24	57.1	9.40E-14	13.0269
PF13894	C2H2-type zinc finger	149	172	23	45.2	3.80E-10	9.4202
PF13894	C2H2-type zinc finger	177	202	25	45.2	3.80E-10	9.4202
PF13465	Zinc-finger double domain	163	190	27	35.8	2.50E-07	6.6021
PF12874	Zinc-finger of C2H2 type	149	171	22	22.2	0.00308	2.5117

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost

	PF00096	PF00096	PF13894	PF13894	PF13465	PF12874
PF00096	-1	-1	0	-1	1	1
PF00096	-1	-1	-1	0	1	-1
PF13894	1	-1	-1	-1	0	1
PF13894	-1	1	-1	-1	0	-1
PF13465	0	0	1	1	-1	1
PF12874	0	-1	0	-1	0	-1

a. Proposed domains in FGSG_16768

PF13894 (149-172)**, PF13894** (177-202)

b. Justification

There are two sets of overlapping domains:

- Set one: PF00096 (149-171), PF13894 (149-172), PF12874 (149-171)
- Set two: PF00096 (177-201), PF13894 (177-202)

Moreover, domain PF13465 (163-190) overlaps with all domains from both sets.

Solving the overlapping in set one, including domain PF13465 (163-190):

Domains PF12874 (149-171) and PF13465 (163-190) were excluded based on the score (Table D-7 B). The score for domains PF00096 (149-171) and PF13894 (149-172) is the highest and equal to each other. Thus, considering the pairwise score comparison for these two domains, domain PF13894 (149-172) was chosen.

Solving the overlapping in set two, including domain PF13465 (163-190):

As previously domain PF13465 (163-190) was excluded; it is not considered in solving the overlapping in set two. The score for both domains PF00096 (177-201) and PF13894 (177-202) are one of the highest and equal to each other. Thus, considering the pairwise score comparison of these two domains, domain PF13894 (177-202) was chosen.

7. FGSG_01410 probable myosin I heavy chain

Table D-8 Overlapping domains in FGSG_01410

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF00063	Myosin head (motor domain)	40	699	659	780.85	1.30E-231	230.8861
PF06017	Myosin tail	756	959	203	145.5	2.40E-40	39.6198
PF00018	SH3 domain	1075	1122	47	56.66	1.30E-13	12.8861
PF14604	Variant SH3 domain	1076	1126	50	51.48	4.70E-12	11.3279
PF07653	Variant SH3 domain	1073	1128	55	35.38	3.30E-07	6.4815

B. Score table generated by python script (-1) does not overlap, (1 overlap but win, (0) overlap but lost

	PF00063	PF06017	PF00018	PF14604	PF07653
PF00063	-1	-1	-1	-1	-1
PF06017	-1	-1	-1	-1	-1
PF00018	-1	-1	-1	0	1
PF14604	-1	-1	1	-1	0
PF07653	-1	-1	0	1	-1

a. Proposed domains in FGSG_01410:

PF00063 (40-699), **PF06017** (756-959), **PF00018** (1075-1122)

b. Justification

Here two domains do not overlap: PF00063 (40-699) and PF06017 (756-959), and set of three domains that overlap with each other: PF00018 (1075-1122), PF14604 (1076-1126), PF07653 (1073-1128).

All overlapping domains have the same score (Table D-8 B) and applying pairwise score comparison did not resolve the overlapping. Thus, based on the best e-value domain PF00018 (1075-1122) was chosen.

8. FGSG_04993 conserved hypothetical protein

Table D-9 Overlapping domains in FGSG_04993

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF00583	Acetyltransferase (GNAT) family	64	185	121	46.43	1.60E-10	9.7959
PF13508	Acetyltransferase (GNAT) domain	57	186	129	35.88	2.30E-07	6.6383
PF13673	Acetyltransferase (GNAT) domain	15	184	169	27.09	0.000104	3.983

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost

	PF00583	PF13508	PF13673
PF00583	-1	0	1
PF13508	1	-1	0
PF13673	0	1	-1

a. Proposed domain in FGSG_04993

PF00583 (64-185)

b. Justification

Here we have a situation that all domains overlap with each other and neither score nor pairwise scoring for domains (Table D-9 B) solves the problem of the overlapping. Thus, domain PF00583 (64-185) was chosen based on the best e-value.

9. FGSG_04675 conserved hypothetical protein

Table D-10 Overlapping domains in FGSG_04675

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF08241	Methyltransferase domain	52	148	96	56.15	1.90E-13	12.7212
PF13649	Methyltransferase domain	51	144	93	46.97	1.10E-10	9.9586
PF13847	Methyltransferase domain	45	196	151	43.94	8.80E-10	9.0555
PF08242	Methyltransferase domain	52	146	94	40.6	8.90E-09	8.0506
PF13489	Methyltransferase domain	24	202	178	31.48	0.000005	5.301

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost

	PF08241	PF13649	PF13847	PF08242	PF13489
PF08241	-1	1	0	1	1
PF13649	0	-1	0	0	0
PF13847	1	1	-1	1	0
PF08242	0	1	0	-1	0
PF13489	0	1	1	1	-1

a. Proposed domain in FGSG_04675

PF08241 (52-148)

b. Justification

Here we have a situation where all domains overlap with each other. Domains PF08241 (52-148), PF13847 (45-196), and PF13489 (24-202) have the best and equal scores (Table D-10 B). Score pairwise comparison did not either resolve the overlapping problem. Thus, based on the best e-value domain, PF08241 (52-148) was chosen.

10. FGSG_10048 probable RVS167 protein

Table D-11 Overlapping domains in FGSG_10048

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF03114	BAR domain	6	247	241	240.15	7.60E-69	68.1192
PF00018	SH3 domain	380	427	47	59.74	1.50E-14	13.8239
PF14604	Variant SH3 domain	381	431	50	55.91	2.20E-13	12.6576
PF07653	Variant SH3 domain	378	433	55	42.92	1.80E-09	8.7447

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost

	PF03114	PF00018	PF14604	PF07653
PF03114	-1	-1	-1	-1
PF00018	-1	-1	0	1
PF14604	-1	1	-1	0
PF07653	-1	0	1	-1

a. Proposed domains in FGSG_10048:

PF03114 (6-247), **PF00018** (380-427)

b. Justification

The first domain PF03114 (6-247) does not overlap with the rest of domains in the protein, while domains PF00018 (380-427), PF14604 (381-431), and PF07653 (378-433) overlap with each other. Neither score nor score pairwise comparison resolves the overlapping within these three domains. Thus, domain PF00018 (380-427) was chosen based on the best e-value.

11. FGSG_10647 conserved hypothetical protein

Table D-12 Overlapping domains in FGSG_10647

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF08241	Methyltransferase domain	47	144	97	67.51	7.10E-17	16.1487
PF12847	Methyltransferase domain	42	147	105	54.17	7.30E-13	12.1367
PF08242	Methyltransferase domain	47	142	95	47.1	9.80E-11	10.0088
PF13847	Methyltransferase domain	40	225	185	45.78	2.50E-10	9.6021
PF13489	Methyltransferase domain	23	214	191	44.21	7.30E-10	9.1367
PF13649	Methyltransferase domain	46	140	94	32.93	0.000002	5.699

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost

	PF08241	PF12847	PF08242	PF13847	PF13489	PF13649
PF08241	-1	0	1	1	1	1
PF12847	1	-1	1	0	0	1
PF08242	0	0	-1	0	0	1
PF13847	0	1	1	-1	0	1
PF13489	0	1	1	1	-1	1
PF13649	0	0	0	0	0	-1

a. Proposed domain in FGSG_10647:

PF08241 (47-144)

b. Justification

Here all domains overlap with each other. Based on the score (Table D-12 B), domains PF08241 (47-144) and PF13489 (23-214) were chosen for further consideration. Based on score pairwise comparison domain PF08241 (47-144) was chosen.

12. FGSG_08624 conserved hypothetical protein

Table D-13 Overlapping domains in FGSG_08624

A. Domains within the protein (based on hmmer output)

Pfam ID	Domain description	Start	End	End-Start	Score	e-value	-Log10 (e-value)
PF08241	Methyltransferase domain	70	163	93	64.32	6.40E-16	15.1938
PF12847	Methyltransferase domain	65	166	101	52.14	3.00E-12	11.5229
PF08242	Methyltransferase domain	70	161	91	48.94	2.70E-11	10.5686
PF13649	Methyltransferase domain	69	159	90	46.26	1.80E-10	9.7447
PF13847	Methyltransferase domain	63	203	140	45.07	4.00E-10	9.3979
PF13489	Methyltransferase domain	33	209	176	32.45	0.000003	5.5229

B. Score table generated by python script (-1) does not overlap, (1) overlap but win, (0) overlap but lost

	PF08241	PF12847	PF08242	PF13649	PF13847	PF13489
PF08241	-1	0	1	1	1	1
PF12847	1	-1	1	1	0	1
PF08242	0	0	-1	1	0	1
PF13649	0	0	0	-1	0	0
PF13847	0	1	1	1	-1	0
PF13489	0	0	0	1	1	-1

a. Proposed domain for FGSG_08624

PF12847 (65-166)

b. Justification

Here all domains overlap with each other. However, two domains PF08241 (70-163) and PF12847 (65-166) have the highest equal score (Table D-13 B). Pairwise score comparison of these two domains revealed that the best domain is PF12847 (65-166).

IV. Summary

Table D-14 Summary of resolving the domains overlapping issue within 12 FG proteins
Column 4 consists of all domains generated by hmmer version 3 for listed FG protein. The best domains taken into further analysis (stays in FG protein) are highlighted in bold.
This data is relevant to section 5.3.1 of Chapter 5.

No	FG ID	Protein description	Domains (Start-End)	Domain description
1	**FGSG_04675**	**Conserved hypothetical protein**	PF13489 (24-202)	Methyltransferase domain
			PF13847 (45-196)	Methyltransferase domain
			PF13649 (51-144)	Methyltransferase domain
			PF08242 (52-146)	Methyltransferase domain
			PF08241 (52-148)	Methyltransferase domain
2	**FGSG_06199**	**Hypothetical protein**	PF13489 (128-349)	Methyltransferase domain
			PF13847 (149-317)	Methyltransferase domain
			PF12847 (151-253)	Methyltransferase domain
			PF13649 (155-246)	Methyltransferase domain
			PF08242 (156-248)	Methyltransferase domain
			PF08241 (156-250)	Methyltransferase domain
3	**FGSG_08624**	**Conserved hypothetical protein**	PF13489 (33-209)	Methyltransferase domain
			PF13847 (63-203)	Methyltransferase domain
			PF12847 (65-166)	Methyltransferase domain
			PF13649 (69-159)	Methyltransferase domain
			PF08242 (70-161)	Methyltransferase domain
			PF08241 (70-163)	Methyltransferase domain
4	**FGSG_10647**	**Conserved hypothetical protein**	PF13489 (23-214)	Methyltransferase domain
			PF13847 (40-225)	Methyltransferase domain
			PF12847 (42-147)	Methyltransferase domain
			PF13649 (46-140)	Methyltransferase domain
			PF08242 (47-142)	Methyltransferase domain
			PF08241 (47-144)	Methyltransferase domain
5	**FGSG_15775**	**Hypothetical protein**	PF13489 (8-202)	Methyltransferase domain
			PF13847 (33-196)	Methyltransferase domain
			PF12847 (35-149)	Methyltransferase domain
			PF13649 (39-142)	Methyltransferase domain
			PF08242 (40-144)	Methyltransferase domain
			PF08241 (40-146)	Methyltransferase domain
6	**FGSG_16446**	**Related to transcription factor Pig1p**	PF00096 (10-32)	Zinc finger, C2H2 type
			PF13894 (10-33)	C2H2-type zinc finger
			PF13465 (24-49)	Zinc-finger double domain
			PF13894 (38-61)	C2H2-type zinc finger
			PF00096 (38-60)	Zinc finger, C2H2 type
			PF00172 (82-119)	Fungal Zn(2)-Cys(6) binuclear cluster domain
			PF04082 (360-621)	Fungal-specific transcription factor domain

Table D-14 continues

No	FG ID	Protein description	Domains (Start-End)	Domain description
7	**FGSG_16504**	**Hypothetical protein**	**PF13923 (72-124)**	Zinc finger, C3HC4 type (RING finger)
			PF13920 (245-293)	Zinc finger, C3HC4 type (RING finger)
			PF13639 (247-287)	Ring finger domain
			PF14634 (248-288)	zinc-RING finger domain
			PF13445 (249-284)	RING-type zinc-finger
			PF13923 (249-286)	Zinc finger, C3HC4 type (RING finger)
			PF00097 (249-286)	Zinc finger, C3HC4 type (RING finger)
			PF15227 (249-286)	zinc finger of C3HC4-type, RING
			PF02190 (340-583)	ATP-dependent protease La (LON) domain
8	**FGSG_16768**	**Related to krueppel protein**	PF00096(149-171)	Zinc finger, C2H2 type
			PF12874 (149-171)	Zinc-finger of C2H2 type
			PF13894 (149-172)	C2H2-type zinc finger
			PF13465 (163-190)	Zinc-finger double domain
			PF00096 (177-201)	Zinc finger, C2H2 type
			PF13894 (177-202)	C2H2-type zinc finger
9	**FGSG_16806**	**Hypothetical protein**	**PF13920 (12-78)**	Zinc finger, C3HC4 type (RING finger)
			PF13639 (14-72)	Ring finger domain
			PF14634 (15-73)	zinc-RING finger domain
			PF13445 (16-69)	RING-type zinc-finger
			PF00097 (16-71)	Zinc finger, C3HC4 type (RING finger)
			PF13923 (16-71)	Zinc finger, C3HC4 type (RING finger)
			PF00018 (878-925)	SH3 domain
			PF14604 (879-929)	Variant SH3 domain
10	**FGSG_01410**	**Probable myosin I heavy chain**	**PF00063 (40-699)**	Myosin head (motor domain)
			PF06017 (756-959)	Myosin tail
			PF07653 (1073-1128)	Variant SH3 domain
			PF00018 (1075-1122)	SH3 domain
			PF14604 (1076-1126)	Variant SH3 domain
11	**FGSG_10048**	**Probable RVS167 protein**	**PF03114 (6-247)**	BAR domain
			PF07653 (378-433)	Variant SH3 domain
			PF00018 (380-427)	SH3 domain
			PF14604 (381-431)	Variant SH3 domain
12	**FGSG_04993**	**Conserved hypothetical protein**	PF13673 (15-184)	Acetyltransferase (GNAT) domain
			PF13508 (57-186)	Acetyltransferase (GNAT) domain
			PF00583 (64-185)	Acetyltransferase (GNAT) family

Appendix E.

Additional tables to Chapter 6

Table E-1 DUFs identified in *F. graminearum* and *F. venenatum* only.
This data is relevant to Chapter 6, section 6.3.3 (Distribution of all DUFs within tested *Fusarium* proteomes).

Species	DUF Id	Pfam Id	Protein	Protein description	Other domains	NCBI blastp	Cover [%]	Identity[%]
F.graminearum (FG3 MIPS gene call)	DUF1242	PF06842	FGSG_05210	CHP	No	*Nectria haematococca* mpVI 77-13-4	100	97
F.graminearum (FGRRE gene call)						*F. oxysporum* f. sp. *cubense race* 4	91	100
F. venenatum						*Metarhizium robertsii*	94	97
			FGRRES_05210		No	*F. graminearum* PH-1	100	98
						F. oxysporum f. sp. *cubense race 4*	91	98
						Nectria haematococca mpVI 77-13-4	94	96
			FV_09062		No	*Nectria haematococca* mpVI 77-13-4	100	97
						F. oxysporum f. sp. *cubense race 4*	91	100
						Metarhizium robertsii	94	97
	DUF1620	PF07774	FGSG_00261	CHP	PF13360|DUF1620	*F. pseudograminearum* CS3096	100	99
						F. avenaceum	98	90
			FGRRES_00261		PF13360|DUF1620	*F. pseudograminearum* CS3096	100	99
						F. avenaceum	98	90
			FV_00390		PF13360|DUF1620	*F. graminearum* PH-1	100	96
						F. pseudograminearum CS3096	98	96
						F. avenaceum	98	91
	DUF2052	PF09747	FGSG_08947	CHP		*F. pseudograminearum* CS3096	100	86
						F. fujikuroi IMI 58289	100	75
			FGRRES_08947			*F. graminearum* PH-1	100	90
						F. pseudograminearum CS3096	100	86
						F. fujikuroi IMI 58289	100	75
			FV_05637			*F. pseudograminearum* CS3096	100	89
						F. avenaceum	100	83
	DUF2462	PF09495	FGSG_15041	CHP	No	*F. pseudograminearum* CS3096	100	98
						F. acuminatum CS5907	100	79
						F. avenaceum	100	77
			FGRRES_15041			*F. pseudograminearum* CS3096	100	98
						Fusarium acuminatum CS5907	100	79
						F. avenaceum	100	77
			FV_03635		No	*F. graminearum*	100	88
						F. pseudograminearum CS3096	100	87

Where CHP – Conserved Hypothetical Protein, HP - Hypothetical Protein, DUF – Domain of Unknown Function

Table E-2: DUFs identified only in *F. culmorum* and *F. venenatum* proteomes.
This data is relevant to Chapter 6, section 6.3.3 (Distribution of all DUFs within tested *Fusarium* proteomes).

Species	DUF Id	Pfam Id	Protein	Other domains	NCBI blastp	Cover [%]	Identity[%]
F. culmorum	DUF814	PF05670	FCUL_04912	PF05833|DUF814|DUF3441	*F. graminearum* PH-1 (CHP FGSG_08738) - PF05833|DUF3441	100	98
F. venenatum					*F. graminearum* (HP FG05_08738)	100	98
					F. pseudograminearum CS3096 (HP FPSE_07722)	100	97
			FCUL_10279	No	*F. oxysporum* f. sp. *cubense race* 4 (Coiled-coil domain-containing protein 25)	100	94
					F. oxysporum Fo5176 (HP FOXB_10908)	100	94
					F. pseudograminearum CS3096 (HP FPSE_01683)	100	92
			FV_11268	No	*Fusarium* sp. FIESC_5 CS3069 (unnamed protein product)	100	93
					F. fujikuroi (Uncharacterized protein LW93_1646)	100	93
					F. oxysporum f. sp. *cubense race* 4 (Coiled-coil domain-containing protein 25)	100	93
	DUF3435	PF11917	FCUL_13493	No	**No hit**		
			FCUL_09938	No	*F. oxysporum* Fo5176 (HP FOXB_06467)	93	93
					F. oxysporum f. sp. *vasinfectum* 25433 (HP FOTG_17373)	95	92
					F. oxysporum Fo47 (HP FOZG_04448)	95	90
			FV_04335	No	*F. oxysporum* Fo5176 (HP FOXB_06467)	97	85
					F. oxysporum f. sp. *vasinfectu* m 25433 (HP FOTG_17373)	96	85
					F. oxysporum FOSC 3-a (HP FOYG_02304)	96	85
	DUF3505	PF12013	FCUL_13218	DUF3505|DUF3505	*F. oxysporum* f. sp. *radicis-lycopersici* 26381 (HP FOCG_17962)	43	86
			FCUL_13198	DUF3505|DUF3505	*F. oxysporum* FOSC 3-a (HP FOYG_17331)	99	92
					F. oxysporum f. sp. *melonis* 26406 (HP FOMG_17381)	99	91
					F. oxysporum f. sp. *lycopersici* MN25 (HP FOWG_16793)	99	90
			FCUL_13213	DUF3505|DUF3505	*F. oxysporum* f. sp. *vasinfectum* 25433 (HP FOTG_16532)	95	97
					F. oxysporum f. sp. *pisi* (HP FOVG_17233)	95	97
			FCUL_13616	No	*F. oxysporum* f. sp. vasinfectum 25433 (HP FOTG_17521)	70	67
					F. oxysporum f. sp. *conglutinans race* 2 54008 (HP FOPG_18258)	81	67
			FV_11060	DUF3505|PF00270|PF00271	*F. oxysporum* f. sp. *melonis* 26406 (HP FOMG_19288)	81	86
					F. fujikuroi (Uncharacterized protein Y057_10530)	67	89
			FV_07932	No	*F. oxysporum* f. sp. *vasinfectum* 25433 (HP FOTG_10639)	44	64
	DUF3669	PF12417	FCUL_04796	No	*F. langsethiae (* HP FLAG1_02205)	97	65
					Colletotrichum sublineola (HP CSUB01_04146)	98	49
			FV_05446	No	*F. langsethiae (* HP FLAG1_02205)	99	74
					Verticillium longisporum (HP BN1708_004999)	85	62
					Colletotrichum sublineola (HP CSUB01_04146)	95	48

Where DUF – Domain of Unknown Function

Appendix F.

Publications

(This section is formatted according to the original source and starts from the next page)

OPEN ACCESS Freely available online

PLOS ONE

Network-Based Data Integration for Selecting Candidate Virulence Associated Proteins in the Cereal Infecting Fungus *Fusarium graminearum*

Artem Lysenko[1☯], Martin Urban[3☯], Laura Bennett[2], Sophia Tsoka[2], Elzbieta Janowska-Sejda[1,3], Chris J. Rawlings[1], Kim E. Hammond-Kosack[3*☯], Mansoor Saqi[1☯]

1 Department of Computational and Systems Biology, Rothamsted Research, Harpenden, United Kingdom, **2** Department of Informatics, School of Natural and Mathematical Sciences, Kings College London, Strand, London, United Kingdom, **3** Department of Plant Biology and Crop Science, Rothamsted Research, Harpenden, United Kingdom

Abstract

The identification of virulence genes in plant pathogenic fungi is important for understanding the infection process, host range and for developing control strategies. The analysis of already verified virulence genes in phytopathogenic fungi in the context of integrated functional networks can give clues about the underlying mechanisms and pathways directly or indirectly linked to fungal pathogenicity and can suggest new candidates for further experimental investigation, using a 'guilt by association' approach. Here we study 133 genes in the globally important Ascomycete fungus *Fusarium graminearum* that have been experimentally tested for their involvement in virulence. An integrated network that combines information from gene co-expression, predicted protein-protein interactions and sequence similarity was employed and, using 100 genes known to be required for virulence, we found a total of 215 new proteins potentially associated with virulence of which 29 are annotated as hypothetical proteins. The majority of these potential virulence genes are located in chromosomal regions known to have a low recombination frequency. We have also explored the taxonomic diversity of these candidates and found 25 sequences, which are likely to be fungal specific. We discuss the biological relevance of a few of the potentially novel virulence associated genes in detail. The analysis of already verified virulence genes in phytopathogenic fungi in the context of integrated functional networks can give clues about the underlying mechanisms and pathways directly or indirectly linked to fungal pathogenicity and can suggest new candidates for further experimental investigation, using a 'guilt by association' approach.

Citation: Lysenko A, Urban M, Bennett L, Tsoka S, Janowska-Sejda E, et al. (2013) Network-Based Data Integration for Selecting Candidate Virulence Associated Proteins in the Cereal Infecting Fungus *Fusarium graminearum*. PLoS ONE 8(7): e67926. doi:10.1371/journal.pone.0067926

Editor: Yin-Won Lee, Seoul National University, Republic of Korea

Received December 7, 2012; **Accepted** May 23, 2013; **Published** July 4, 2013

Funding: MS, MU and KHK are supported by the BBSRC (http://www.bbsrc.ac.uk) through the Institute Strategic Programme 20:20 Wheat® (grant BB/J/00426X/1). In addition the PHI-base project receives support from the BBSRC grant BB/I000488/1 and AL was supported by BBSRC grant BB/G015716/1. ST acknowledges support from the Leverhulme Trust (http://www.leverhulme.ac.uk, grant RPG-2012-686). The PHI-base project also receives support as a BBSRC National Capability (BB/J/004383/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kim.hammond-kosack@rothamsted.ac.uk

☯ These authors contributed equally to this work.

Introduction

The Ascomycete fungus *Fusarium graminearum* (teleomorph *Gibberella zeae*) is a major pathogen of wheat causing *Fusarium* ear blight, Fusarium head blight or *Fusarium* head scab disease [1,2] (www.scabusa.org). As wheat accounts for 32% of global cereal production and provides 20% of the world's calorific intake (www.fao.org), control strategies for *Fusarium* infection are important for food security. *Fusarium graminearum* can also infect the floral tissue of numerous other cereal species, including maize, barley, triticale, rice and oats [1]. Although affecting yield, *Fusarium* infection often leads to reduced grain quality and to contamination of the grain with various mycotoxins, in particular the ß-type trichothecene deoxynivalenol (DON) and its acetylated derivatives (15A-DON and 3A-DON), which may make the grain unsafe for human and/or animal consumption [3].

The genome sequence of *Fusarium graminearum* [4] is predicted to code for 13,332 proteins and further revisions to the identification of open reading frames and annotation are in progress [5,6]. As a result of the analysis of a genetic cross between the sequenced strain and another strain, the *F. graminearum* genome is recognised to contain regions of high recombination in both sub-telomeric and central chromosome regions interspersed with longer regions with low or no genetic recombination. Genes shared between different Fusarium species are primarily located in the low and no recombination regions [4]. Particular genes in *F. graminearum*, other Fusarium species and other plant fungal pathogens have been investigated experimentally for their contribution to pathogenicity or virulence, i.e. their qualitative or quantitative effect of the disease causing ability of a microbe. Typically these experiments involve stable gene disruption/gene deletion in the pathogen and observation of the resulting infection phenotype in one or more host plant systems. Already a large number of *F. graminearum* genes have been tested and published, of which 100 were found to alter virulence and 33 had no effect on the interaction tested at the time

of writing this article [7] and **Table S1**). Several of these *F. graminearum* virulence genes are unique to this species or restricted to closely related Fusarium species whilst others genes are also required for virulence in other plant and/or animal infecting microbes. To assist comparative studies, the functions in numerous other pathosystems of pathogenicity and virulence associated genes has been catalogued in the Pathogen-Host Interactions database called PHI-base [8–10], accessible at www.phi-base.org. This is an expertly curated database for ~1000 pathogen-host interactions. The plant, animal, fungal, oomycete and/or bacteria entries in PHI-base are extracted from the scientific literature by domain experts and therefore describe experimentally tested interactions, for example the effect of a given gene disruption experiment in a given pathogen, on a particular host. Importantly PHI-base also details those tested genes, which had no effect on pathogenicity.

In order to understand how particular genes and their gene products may contribute to the pathogenic process it is necessary to explore the biological context of these genes. Approaches that involve placing these genes within various relationship networks provide a useful starting point. The relationships can include, for example, gene co-expression, known or predicted protein-protein interactions, and sequence similarity (see for example [11]). Previously, a predicted protein-protein interaction (PPI) network has been used to predict pathogenicity genes in *Fusarium graminearum* [12]. This 'guilt by association' approach [13] was used to examine those proteins in a predicted PPI network [14] that have at least two known pathogen associated genes as nearest neighbours with additional filtering of candidates using some of the available *in planta* and *in vitro* gene expression data available from a comprehensive data source called PLEXdb [15]. The Liu et al. network analysis used an initial list of 49 *F. graminearum* gene sequences available in PHI-base. A total of 39 potential virulence associated proteins were identified, of which nine have now been connected to virulence through experimentation (reviewed in [7]).

Here we extend the study of [12] by using an integrated network that includes co-expression information and sequence similarity in addition to the core predicted PPI network [14] as well as a larger set of known *Fusarium graminearum* virulence associated genes. The aim of this study was two-fold: firstly to predict additional *F. graminearum* virulence associated genes that could then become targets for experimental analysis and secondly to enable the biological context of the predictions to be explored. As our starting point, we have used the set of verified virulence (VV) genes taken from the pathogen host interaction database PHI-base (version 3.3) as well as manual curation of the recent literature on *Fusarium graminearum* pathogenicity in order to include entries not yet in PHI-base 3.3. The data integration has been carried out using the Ondex data integration and visualisation system [16,17] which allows the integrated network to be explored manually. The filtering tools in the Ondex system allow the effects of inclusion or exclusion of various evidence types on the predictions to be inspected. We discuss in detail the biological plausibility of some of the predictions. The predictions in the context of the entire network have been made available for use by the community. We acknowledge that the term virulence associated genes/proteins can be interpreted in a number of ways – the candidates we have identified may be involved in some part of the virulence process but not necessarily be directly involved (for example, an effector protein) and could be seen as system components [18].

Results

Predictions Made with the Integrated Network

We constructed an integrated network for *Fusarium graminearum* using information from protein sequence similarity, gene co-expression and predicted protein interactions (PPI). The coexpression links were created between nodes representing proteins if the genes encoding them were found to be coexpressed. We have previously described the disjoint and overlapping community structure of the integrated network in [19]. Here we use the network for prediction of potential new virulence associated proteins.

Table 1 shows the graph topological properties, calculated with the NetworkX package [20], of the three constituent networks, as well as an integrated network, which uses information from all three constituent networks. The sequence similarity network has a large number of connected components (subgraphs in which any two nodes are connected by a path of edges) and a high transitivity measure (suggesting more tightly connected structures, i.e a more 'clique-like' structure). These properties most likely reflect the grouping of the proteins into sequence similar groups. The predicted protein interaction network from [14] also has a high transitivity suggesting a more 'clique like' structure, which may be an indication of predicted protein complexes, although this structure may be affected by the way in which data from some experiments is interpreted and represented as binary interactions in different PPI data sources [17].

The prediction of virulence associated proteins was carried out in the Ondex software by the implementation of a new plug-in, as described in the Methods section. Following [12] a node in the network was labelled as a predicted virulence associated protein if it was a nearest neighbour to at least two VV seeds. Fewer VV seeds were involved in predictions in the co-expression network than the PPI network (18 seeds as compared to 30). As expected, the integrated network was the largest and had the greatest number of VV seeds which were involved in predictions (60).

This approach resulted in 215 predictions in the integrated network, which was considerably more than could be predicted from any of the individual constituent networks: using only the sequence similarity based network leads to 100 predictions, the coexpression network yields 47 predictions and the predicted PPI network of Zhao et al (2009) 79 predictions. The 215 predictions (**Table S2**) contain 29 proteins annotated as hypothetical protein in the *Fusarium graminearum* database [6]. The predictions made on the basis of PPI links to the VV seeds may reflect an ancient species conserved sub network, because the *Fusarium* PPI network originally described by [14] had been developed using information from six eukaryotic species and one prokaryotic species, which are all non-pathogenic, namely, budding and fission yeasts, human, mouse, fly, worm and *E. coli*. The predictions made on the basis of co-expression links could potentially represent, either a fungal taxon restricted, but conserved network, a *Fusarium graminearum* specific network or again be part of an ancient species conserved network. The complete list of all predictions and the seeds they are connected to is available as **Table S3**. We have also included what proportion of all edges for each of the predicted nodes are linked to seeds. Although, it would be reasonable to assume that a higher proportion would indicate a more certain prediction, the small numbers of available seeds did not allow us to explore this further as part of this study.

Some predictions were made on the basis of the node being a nearest neighbour to a larger number of VV seeds and these may represent more confident predictions. **Table 2** shows the distribution of the number of seeds to which each predicted

Table 1. Comparison of the global properties of the four predicted networks.

Network type	Nodes	Edges	VV seeds leading to predictions	Connected components (CC)	Size of largest CC	Transitivity	Predictions
Sequence network	6349	27807	19 (12)	1155	625	0.69	100 (61)
Core PPI [**14**]	3459	24348	30 (21)	111	2995	0.85	79 (54)
Co-expression	3654	33272	18 (13)	159	3239	0.42	47 (14)
Integrated	9521	80997	60 (50)	439	8364	0.52	215 (120)

Global properties of the three constituent networks and the integrated network. The sequence similarity network excludes nodes with no edges ('orphan' proteins with no sequence similarity matches). Column 4 is the number of verified virulence (VV) seeds involved in the predictions, using the rule that a node must be connected to at least 2 seeds to be a prediction (in brackets are the corresponding numbers if we require connection to at least 3 seeds); Column 8 gives the number of predictions (in brackets are the corresponding prediction counts if we require a more stringent rule i. e. a node must be connected to at least 3 seeds to be a prediction).
doi:10.1371/journal.pone.0067926.t001

virulence associated node is connected in the integrated network as well as in the three constituent networks.

The method for selection of candidate virulence associated proteins based on the network neighbourhood of the proteins previously reported to be important for infection and disease formation was reported by the study of [12]. However, the original study did not validate the underlying assumption that proteins important for virulence are in fact more likely to be connected to other proteins with similar properties. To test this assumption, the node labels were permuted 10000 times to give an estimate of how likely any protein annotated to be involved in virulence is to be connected with at least two others. As shown in **Table 3**, we have observed that the probability is significantly higher than would be expected by chance for sequence similarity and protein-protein interaction networks, but not so for the co-expression network. This result can be taken as an indication that the selection strategy used in this work can be used to reveal the most relevant candidate proteins.

We have described the community structure of the largest connected component of the integrated network in another study [19]. First a series of disjoint (non-overlapping) communities of the network were detected using the Louvain method [21] (which optimises a measure known as modularity [22]. Modularity optimisation is a widely accepted method for community structure detection and has proven its utility in many biological applications and in particular has found functionally coherent communities in PPI networks [23,24]. These disjoint communities were then transformed into overlapping communities through the application of a mathematical programming method, which allows nodes making connections across community borders to be multi-clustered according to the optimisation of another metric known as community strength [19]. In the transformation from disjoint to overlapping communities, the extent of overlapping, i.e. the number of proteins that belong to multiple communities, is controlled by a parameter r. In general, the multi-clustered proteins were found to have a higher connectivity and higher multi-functionality based on Gene Ontology (GO) annotations than proteins belonging to only one module. We found that overall the verified virulence proteins did not appear to show a tendency to belong to multiple communities although one case was noted ($r = 0.4$), where nearly half (49.3%) of the VV proteins belonged to more than one community. We are aware that the small number of VV proteins makes it difficult to ascribe biological significance to these results. We explore here whether the 215 predictions also exhibit the same behaviour. We find that 164 out of the 215 are in the largest module (of size 1951 nodes), which also contains 33 seeds. This module was previously shown to be significantly enriched for VV proteins, and therefore, it makes sense that a large number of the predictions also belong to this community due the nature of guilt-by-association. We now consider the module membership of the predictions to determine whether they tend to belong to more than one module. We find that according to the Fisher's exact test, a significant proportion of the predicted proteins do belong to more than one module (in the range $0.4 \leq r \leq 0.9$). This may be due to the fact that multi-clustered proteins tend to be more connected than proteins belonging to only one module and therefore have a higher chance of being connected to the VV proteins. However, it may also indicate that the proteins

Table 2. Predicting virulence nodes based on the seed numbers connected within the local neighbourhood.

	Number of nodes connected to a given number of seeds			
Number of seeds	**Integrated**	**Protein-protein interaction**	**Co-expression**	**Sequence similarity**
2	95	25	33	39
3	58	48	11	23
4	32	6	3	25
5	23	0	0	12
6	3	0	0	1
7	3	0	0	0
8	1	0	0	0

The number of seeds to which each predicted virulence node is connected, in the four networks is shown. A node linked to 2 or more seed nodes is termed a prediction. Some predictions have links to multiple seeds.
doi:10.1371/journal.pone.0067926.t002

Table 3. The probability that a verified virulence (VV)seed is connected to at least 2 others by chance.

Network type	Seeds connected to 2 or more other seeds	p-value
Integrated	13	0.0001
Protein-protein interaction	4	0.0186
Co-expression	3	0.1172
Sequence similarity	7	7.00E-04

doi:10.1371/journal.pone.0067926.t003

predicted to be virulence associated may have a tendancy to be multi-functional.

We also compared the length distributions of the set of 215 predicted virulence associated proteins with the length distribution of all the proteins in the *F graminearum* genome and find that average lengths of proteins in the predictions and the seeds subset are significantly greater than all the other *F. graminearum* predicted proteins ($n = 12{,}984$) (Student's t-test, $t = 4.49$, $d.f. = 79.30$, $p<0.01$ for seeds vs. other and $t = 4.03$, $d.f. = 225.48$, $p<0.01$ for predictions vs. other). The larger mean size of the VV seeds compared to the 'others' category has arisen purely as a result of the initial protein types selected by the global fusarium community for functional experimentation. The underlying reasons for the increased length of the predicted virulence associated proteins compared to all the other proteins predicted from the *F. graminearum* sequenced genome is currently unclear. However, this analysis clearly indicates that small protein sequences are under-represented in the predictions. The 29 hypothetical proteins predicted have the size range 69 to 1399 amino acids (aa) (mean 527 aa), with only 4 proteins having length under 200 aa (FGSG_01228 (186), FGSG_00536 (116), FGSG_01888 (69), FGSG_08359 (178). We also explored the overall predictive power of the four different networks (**Table S4**). This analysis revealed a marked improvement over the random model. However, the small number of positive and negative examples are insufficient to make an accurate estimate for either the sensitivity or specificity values.

Taxonomic Diversity of the Predictions

The taxonomic diversity of the 215 predicted virulence associated proteins was explored by matching the sequences against the non-redundant database at NCBI (www.ncbi.nlm.nih.gov) so as to obtain an indication of which of the predictions is *Fusarium* or fungal specific. This distribution is represented as a heatmap (**Figure 1**), and the details for each FGSG gene are shown in **Table S5.**

Twenty-five of the predictions are specific up to the level of fungi, whilst 15 are specific up to the level of *Ascomycota*. The FGSG_10808 (a conserved hypothetical protein) and FGSG_03534 (trichothecene 15-O-acetyltransferase) are highly specific to the level of *Hypocreales*. This analysis also highlights that there are 15 predictions unique to the integrated network. Of these six are found to have a taxonomic distribution beyond eukaryotes. Overall, this analysis confirms that the predictions present within each network are for sequences shared with many other eukaryotic species as well as in some case prokaryote species.

Exploring Predictions from Connections to Multiple Seeds

The requirement for a node in the network to be a candidate for virulence was connected to at least two seed VV nodes. As can be seen from **Table 2**, some nodes were connected to a greater number of seeds and these may be suggestive of stronger predictions. One prediction (FGSG_06878) was made on the basis of 8 links to seed proteins. The annotations of the seeds that contributed to the prediction of this protein are given in **Table 4.**

To facilitate the detailed analysis of the network neighbourhood of the predicted virulence associated nodes of interest, the Ondex visualisation tool was used (**Figure 2, Figure S1** for complete neighourhood). These Ondex displays permit the experimenter to explore simultaneously the details associated with each node as well as the origin of the different types of source information via inspection of the colour of each edge connecting the seed to the predicted node.

The prediction FGSG_06878 is linked to 5 seeds with associated phenotype 'reduced virulence', namely FGSG_10313 *(MGV1)*, FGSG_00385, FGSG_08737, FGSG_01964, FGSG_09897 *(SNF1)*, and 3 seeds (FGSG_09903 *(PKAR)* and FGSG_06385 *(FMK1)* and FGSG_16491 *(FST11)* with associated phenotype 'loss of pathogenicity'. This predicted virulence associated protein, FGSG_06878 is annotated in GenRE database [6] as a ''probable CMK1 - Ca2+/calmodulin-dependent ser/thr protein kinase type I''. The prediction and the seeds from which this prediction was inferred are shown in **Figure 2**. Evidence for crosstalk between Map kinase (MAPK) and calcium-calmodulin dependent signalling leading to the activation of transcripton factors was established earlier and was recently reviewed for several plant human pathogenic fungi. [25]. A recent gene deletion study by [26] confirmed a reduced virulence phenotype for FGSG_06878.

The full details of three other predictions that have links to 7 seeds are given in **Table S6** and the immediate networks are displayed in **Figures S2, S3 and S4.** In each case, at least one of the seeds is annotated to be a transcription factor and the prediction is made from information obtained from only two of the constituent networks.

Other Examples of Specific Predictions

In total, this integrated network analysis has predicted 215 potential virulence associated proteins. For illustrative purposes three very different types of predictive example are discussed in detail. The first example was selected because it illustrates the effect of multiple complementary information types contributing to the prediction, the second because a protein unique to *F. graminearum* was predicted and the third example reveals that a network study can identify a specific class of proteins required for virulence, but is unable to pin-point the specific member of a multigene family.

Example 1: Prediction of FGSG_00559 with a Role in Intracellular Signalling Modulation

Within the integrated network, the protein coded for by the gene FGSG_00559 is predicted on the basis of links to four VV

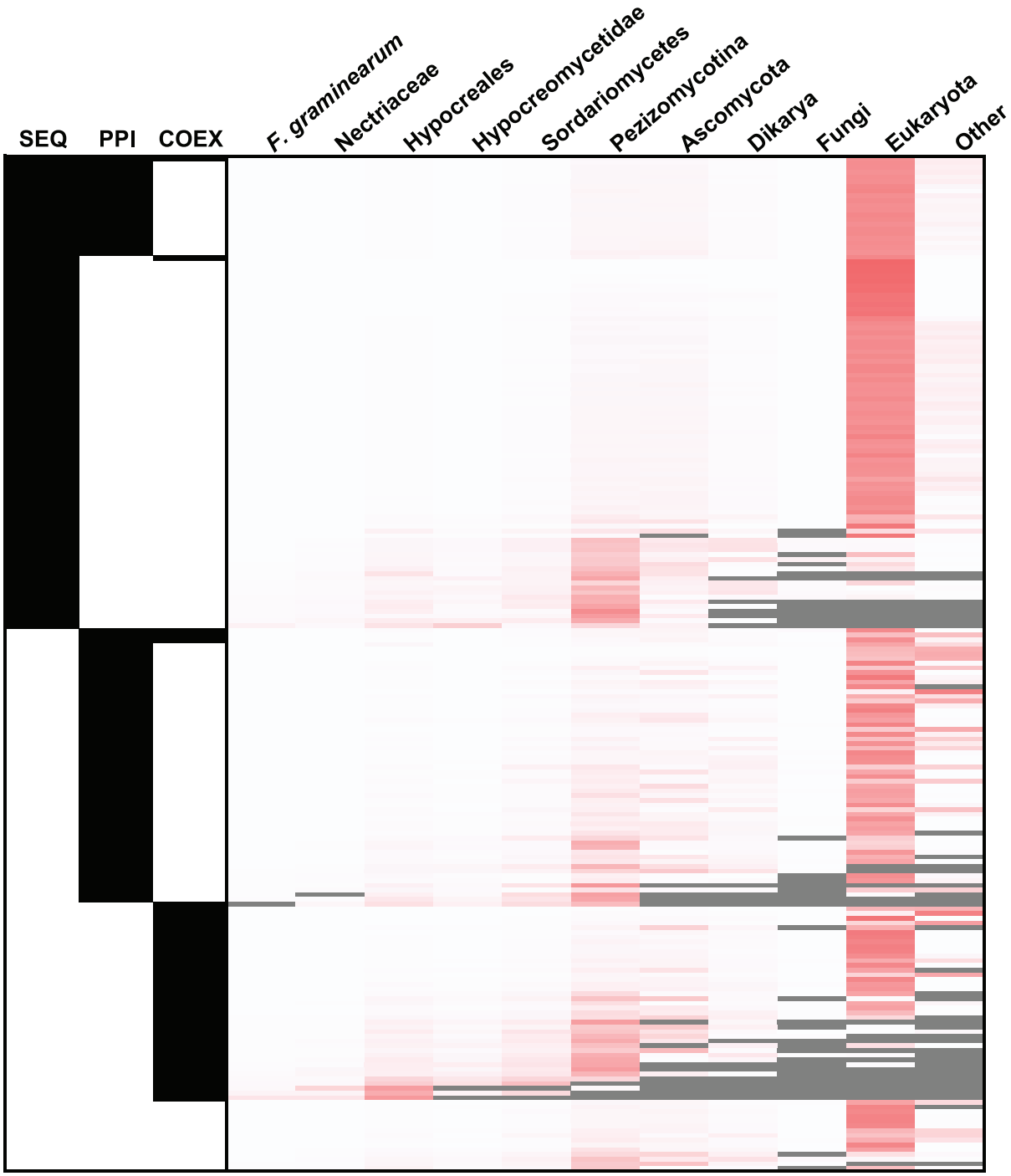


Figure 1. Heat map displaying the taxonomic distribution for each of the predicted virulence associated proteins. Each row provides the information for one sequence. The left hand three columns (SEQ, PPI, COEX) indicate the network in which the prediction could be made (black). For the bottom 15 rows only the integrated network provides the prediction. The right hand heatmap shows the proportional distribution of all BLAST hits from the 215 predictions to the NCBI nr database (white – lowest, red - highest) across the taxonomical levels. All hits were counted once, at the lowest possible level of taxonomical specificity. The grey colour shows cases where there were no hits at a particular taxonomic level. See **Table S2** for the detailed results for each individual FGSG protein.
doi:10.1371/journal.pone.0067926.g001

proteins. Two of these links come from the predicted PPI information (magenta edges in **Figure 3**), namely links to FGSG_06948 (*Gzsc9*, loss of pathogenicity, related to tetratricopeptide repeat protein tpr1) and FGSG_09197 (*HMR1*, reduced virulence, probable 3-hydroxy-3-methylglutaryl-coenzyme A reductase), whilst two links to other proteins are from co-expression

Table 4. The prediction of FGSG_06878 as a virulence factor with links to 8 seeds.

FGSG_06878 (probable CMK1 - Ca2+/ calmodulin-dependent ser/thr protein kinase type I) is linked by	**Seeds on which the prediction is based with phenotype [], and MIPS annotation. Phenotype symbols are rv = reduced virulence, lp = loss of pathogenicity**		
Predicted PPI to:	**FGSG_09903** *(ste7)* [lp], Probable map kinase kinase	**FGSG_10313** [rv] *(mgv1)* (MGV1 map kinase)	**FGSG_06385** *(map1)* [lp] (FMK1 pathogenicity map kinase 1
Co-expression to:	**FGSG_08737** *(GzOB031)* [rv] Probable woronin body major protein precursor	**FGSG_01964** *(CHS5)* [rv] Probable chitin synthase	**FGSG_00385** *(GzHMG002)* [rv] probable NHP6B - nonhistone chromosomal protein
Sequence similarity to:	**FGSG_09897** *(snf1)* [rv] probable serine/ threonine protein kinase	**FGSG_06385** *(map1)* [lp] (FMK1 pathogenicity map kinase 1)	**FGSG_16491** *(fst11)* [lp]] related to NRC-1 MAPKK kinase

This prediction FGSG_06878 was confirmed to be a virulence protein in the recent paper of [21-Wang et al.]. Note that prediction FGSG_06878 is linked to seed FGSG_06385 by both predicted PPI and sequence similarity information. In planta phenotypes are rv, reduced virulence, a quantitative reduction in disease causing ability and the more stringent lp, indicating loss of pathogenicity where disease establishment is aborted.
doi:10.1371/journal.pone.0067926.t004

information (blue edges), namely links to FGSG_09895 (*NTH1*, reduced virulence, probable a neutral trehalase (alpha,alpha-trehalose glucohydrolase)) and FGSG_09908 (*PKAR*, reduced virulence, probable cAMP-dependent protein kinase regulatory chain. FGSG_00559 is annotated in the MIPS GenRE database [6] as a probable 26S proteasome regulatory subunit YTA3. Two of these seed proteins FGSG_09895 and FGSG_09908 reside within close physical proximity in the genome, in a micro-region of virulence genes recently identified using a genome landscape scanning – reverse genetics approach [27,28]. Other predictions included in this network neighbourhood, involving at least two of the same seed proteins include FGSG_06886 a probable 20S core proteasome subunit PRE2, FGSG_09689 a probable ubiquitin-protein ligase (E1-like ubiquitin-activating enzyme) and FGSG_08421 a conserved hypothetical protein. This neighbourhood is highly likely to be involved in co-ordinating two different types of intracellular signalling and possibly involves the degradation of specific signalling components within the proteasome. All the genes in this network neighbourhood were found to reside in regions of either very low or no genetic recombination within the genome [27] and these sequences are found in many fungal and other eukaryotic species (**Table S5**).

Example 2: Prediction of FGSG_00071 Includes Links to Seeds with Opposite Effects

The protein coded for by the gene FGSG_00071 *(TRI1)* is predicted on the basis of links to three VV proteins (**Figure 4**), namely FGSG_16251 (reduced virulence, TRI6, transcription factor) [29], FGSG_03543 (reduced virulence, TRI14, putative trichothecene biosynthesis protein [30] and FGSG_10397 (increase in virulence, CLM1, longiborneol synthetase [31] and FGSG_17598 (recently renamed by MIPS). Previously this gene sequence had been functionally tested as gene FGSG_00007 (increased virulence, cytochrome P450 monooxygenase, DON biosynthesis) [32]. The three other *TRI* genes in this network neighbourhood, namely *TRI3* (FGSG_03534), *TRI4*

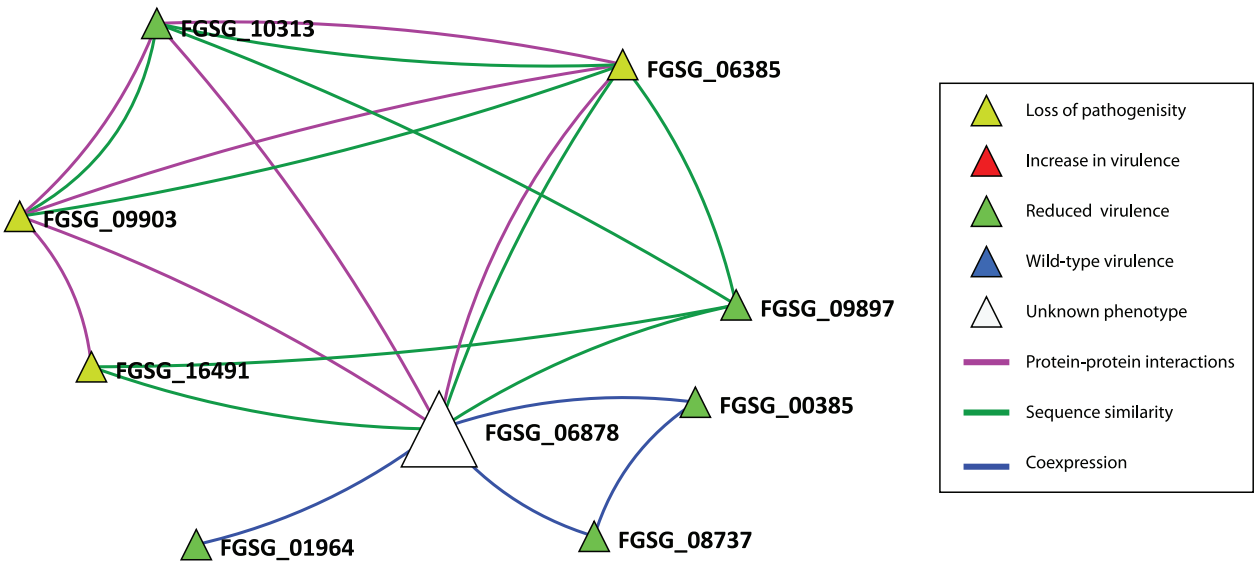


Figure 2. The local neighbourhood for the predicted virulence gene FGSG_06878. The neighbourhood of FGSG_06878 (prediction -large white triangle) and these 8 seed proteins to which it is linked, visualised with Ondex [16]. The magenta coloured edges predicted PPI information, blue edges predicted co-expression information and the green coloured edges predict sequence similarity information.
doi:10.1371/journal.pone.0067926.g002

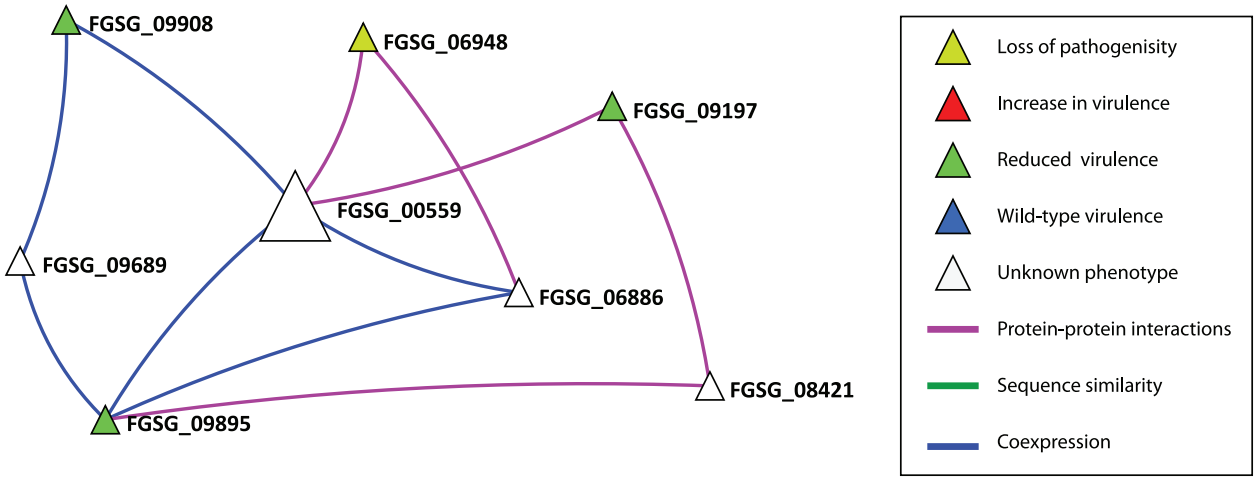


Figure 3. The local neighbourhood for the predicted virulence gene FGSG_00559. The immediate neighbourhood in the integrated network within which the predicted virulence associated protein FGSG_00559 resides (large white triangle). Shown are the types of links between the predictions and the seeds. Magenta coloured edges predicted PPI information and blue edges predicted co-expression information. The various node colours of the seeds as shown in the legend indicate the experimentally determined outcomes. There are 3 additional virulence predictions in this neighbourhood (small white triangles).
doi:10.1371/journal.pone.0067926.g003

(FGSG_03535), TRI11 (FGSG_03540) genes are all located within the main trichothecene (*TRI*) biosynthetic cluster, which is in the middle of chromosome 2 in a region of moderately high genetic recombination. These three *TRI* genes are either suggested or have been shown experimentally in *F. graminearum* to code for key steps in the synthesis of various trichothecene mycotoxins, required for deoxyvalenol (DON) and its acetylated derivatives [33].

TRI1 and FGSG_00007/FGSG_17598 are located towards the left end of Chromosome 1, in the region of very high recombination. FGSG_00007/FGSG_17598 is highly expressed under DON inducing conditions. FGSG_17598 is annotated by

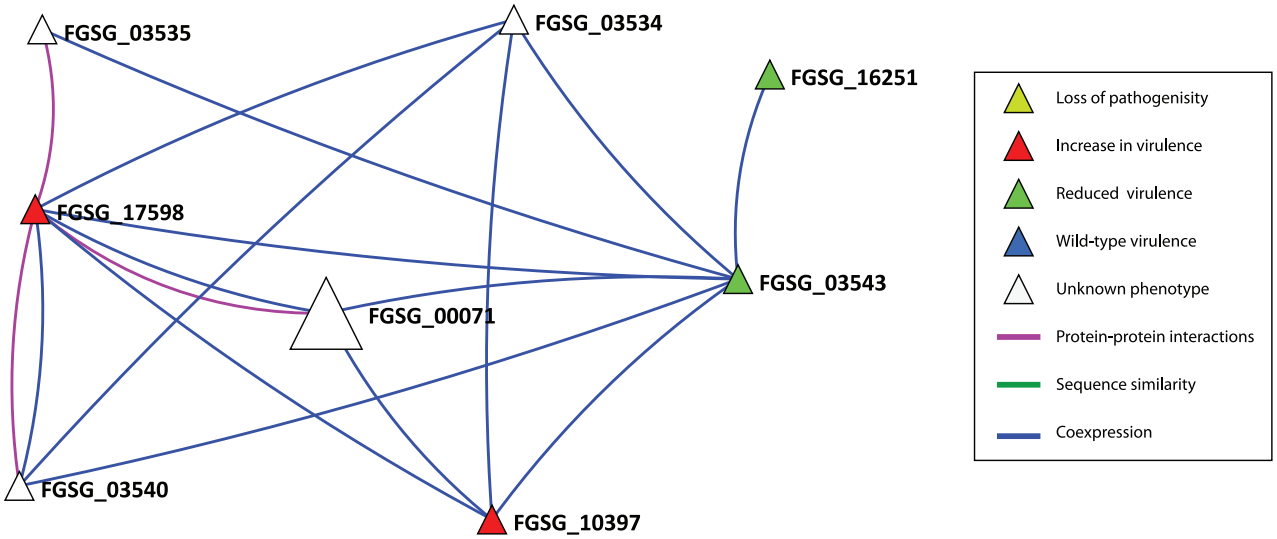


Figure 4. The local neighbourhood for the predicted virulence gene FGSG_00071 *(TRI1).* Gene IDs are: FGSG_03543 *(TRI14)*, FGSG_10397 *(CLM1)*, FGSG_17598 (related to O-methyl sterigmatocystin oxidoreductase), FGSG_03535 *(TRI4)*, FGSG_03534 *(TRI3)*, FGSG_16251 *(TRI6)*, FGSG_03540 *(TRI11)*.
doi:10.1371/journal.pone.0067926.g004

GenRE as 'related to O-methylsterigmatocystin oxidoreductase', but its detailed function is currently unknown.

FGSG_10397 is located in a region of very low recombination at the other end of chromosome 1 and is required for the biosynthesis of a different secondary metabolite, namely Culmorin, when grown under in vitro conditions [31]. In a second study, ([24]-Gardiner et al.) revealed that deletion of FGSG_10397 led to elevated DON mycotoxin production and hence enhanced virulence. However, the level of Culmorin was not reported in the second study.

The predicted virulence node FGSG_00071, is annotated by MIPS as 'TRI1 cytochrome P450 monooxygenase'. A gene disruption mutant in *F. graminearum* was shown to accumulate calonectrin compounds, and no longer produced 15-acetyldeoxynivalenol [34], however the *in planta* phenotype of this mutant strain has not been reported.

Various *TRI* genes are highly expressed during the symptomless phase of wheat ear colonisation when the fungal hyphae are exclusively extracellularly colonising and are in low abundance [35]. This network neighbourhood which contains conflicting experimental results (both enhanced and reduced virulence phenotypes) appears to be involved in both positively and negatively regulating the production of the trichothecene mycotoxin deoyxnivalenol and its acetylated derivatives as well as one other unrelated secondary metabolite, Culmorin, in response to different external stimuli. Most of the *TRI* genes in *F. graminearum* are highly taxon specific**.** This virulence prediction was made on the basis of two co-expression links and one protein interaction link and suggests value in combining multiple data sources (**Figure 4**). The predicted virulence node FGSG_00071 is specific up to the level of *F. graminearum* (**Table S5).**

Example 3: Prediction of Two Non Pathogenicity Associated Seeds as Potential Candidates for Virulence

The two genes FGSG_05535 and FGSG_09988, annotated in GenRE as probable G protein alpha subunits, were shown to be dispensable for pathogenicity [36]. However, both proteins are connected to two seed proteins required for pathogenicity (reduced virulence phenotype). The seed proteins are: FGSG_09614 (GPA2) encoding a guanine nucleotide-binding protein alpha-3 subunit and FGSG_04104 (GPB1) encoding a guanine nucleotide-binding protein beta subunit. Both these seeds are involved in intracellular signalling. The two non-pathogenicity associated proteins as well as 7 others (white triangles in **Figure 5**) would all be predicted to be virulence associated proteins on the basis of having two links to pathogenicity associated seeds. This network neighbourhood contains mostly genes located in genomic regions with very low/no genetic recombination, which are also found in many other taxa. The only exception is FGSG_04618 which is located in a region of very high recombination towards the right hand end of chromosome 2, but which also has a wide taxon distribution. FGSG_09988 codes for the G protein alpha 3 subunit. This reveals the selective recruitment of the G protein alpha subunit to virulence signalling over the beta or gamma subunits in *F. graminearum*. Although this network analysis has revealed a multigene family to be associated with virulence, only through completion of the gene deletion experiments could the actual member recuited to virulence be revealed. None of other members of this cluster belong to multigene families. However the seven other predicted members of this G-protein cluster possess a WD repeat domain.

Example 4: Prediction of Three Non Pathogenicity Associated Seeds as Potential Candidates for Virulence

The gene FGSG_00472 is connected to 5 seeds (**Figure S5**) and is annotated in GenRE as a probable cAMP dependent protein kinase. This gene has recently been shown to be required for pathogenicity and DON production *in planta* [26]. The 5 seed proteins in this cluster are all predicted to be protein kinases. In addition, in this cluster gene FGSG_00472 is connected to two additional potential candidates for virulence, namely genes FGSG_10095 and FGSG_01312. These genes are also annotated in GenRE as protein kinases and are themselves connected to either 3 or 4 protein kinase seeds. Both FGSG_10095 and FGSG_01312 have recently been shown to required for pathogenicity and DON production *in planta* [26]. Interestingly, the three newly verified virulence genes when deleted individually have only a minimal affects on *in vitro* growth, whereas all the seed genes in this cluster when deleted individually have a far greater affect on *in vitro* growth [26].

Mapping of Recently Identified Kinase Proteins in *Fusarium graminearum* to the Integrated Network

The recent comprehensive study of the contribution of the predicted *F. graminearum* kinome to pathogenicity towards wheat ears, mycotoxin production and an additional 15 growth and development traits assessed *in vitro* [26] lead to the identification of 21 putative essential proteins, 44 proteins as having a proven role in disease formation (corresponding to reduced virulence) and 51 proteins with no apparent role in pathogenicity (refer to **Table S7**). We have used this data in an attempt to quantify the predictive accuracy of our combined network approach. Of these 44 new pathogenicity proteins, 23 correspond to predictions made within our integrated network (**Table 5**) and a further 4 are among our set of verified virulence seed proteins (FGSG_10313, FGSG_06385, FGSG_09903, FGSG_09897). In total, 11 of the essential for life proteins in [26] were among our predicted pathogenicity proteins as well as 22, which have been shown to be unaffected in virulence towards wheat ears. This latter figure highlights the problem with false positives. However, some of these single gene negative results may have occured via genetic redundancy, i. e. a member of a multigene family, where the role of the deleted gene can be fully taken over by the function of another related gene(s) and therefore no change in the phenotypic outcome is observed. Only by exploring the effects of deleting specific combinations of sequence related genes can these negative phenotypic effects be confirmed. A further possibility is that some of the predicted virulence genes may only be required for the infection of non-wheat host species.

Selecting only those predictions, which were made on the basis of slightly more stringent criteria, namely requiring at least 3 instead of 2 neighbours as seeds (of which there are 71) has only a small effect with lowering the number of correctly predicted proteins with the phenotype 'reduced virulence' to 21 and with phenotype 'unaffected' to 17.

Chromosomal Location of the Predicted Pathogenicity Associated Proteins

When the newly sequenced *F. graminearum* genome of strain PH-1 and partial sequence information for a second strain GZ3639 were aligned to the available genetic map involving both these strains, this revealed an unanticipated result. Cuomo et al., (2007) described a genome, where the four *F. graminearum* chromosomes were unevenly divided into two types of genomic landscape. The majority of the genome exhibited minimal DNA polymorphism

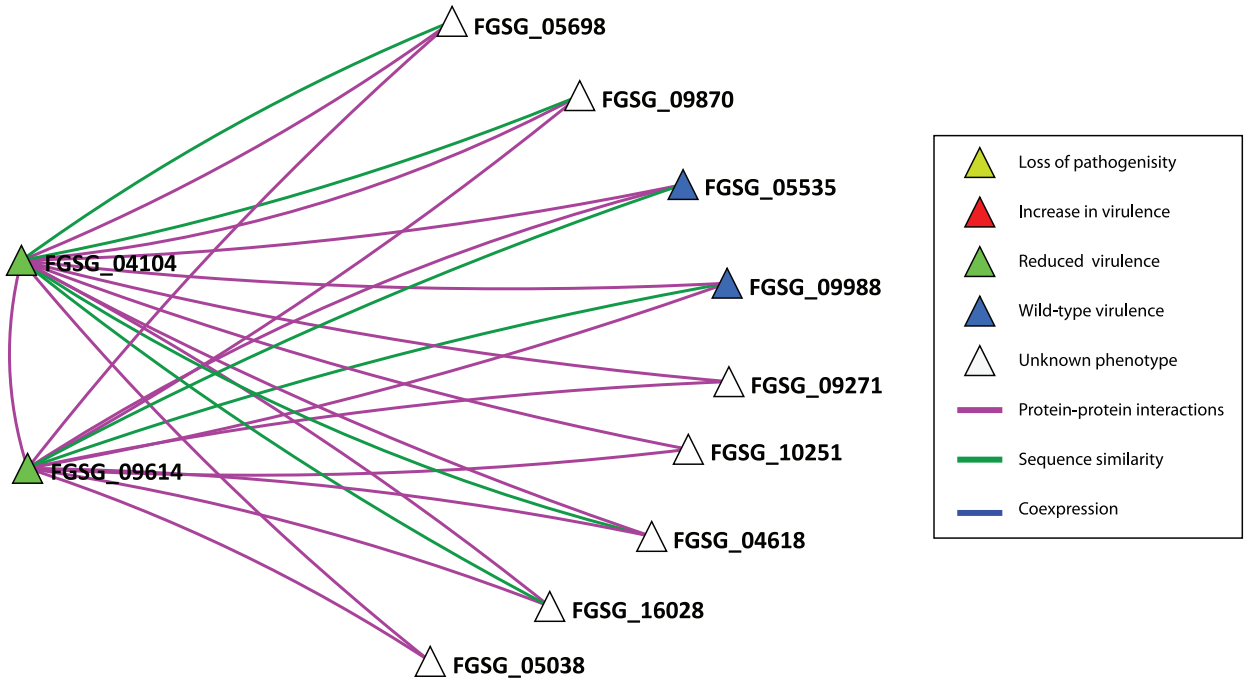


Figure 5. The neighbourhood of FGSG_05535 and FGSG_09988. Although connected to the two seed proteins FGSG_09614 *(GPA2)* and FGSG_04104 *(GPB1),* experimental evidence in barley suggests that the two predictions 05535 and 09988 are dispensable for pathogenicity [36]. Genetic redundancy is suggested to explain this fact. (FGSG_05698: probable *CPC2* protein, FGSG_09870: probable *CPC2* protein; FGSG_09271: probable *SEC13* - protein transport protein; FGSG_10251: probable *LST8* protein; FGSG_04618: related to vegetatible incompatibility protein *HET-E-1*; FGSG_16028: probable U5 snRNP-specific 40 kD protein (novel WD-40 repeat protein); FGSG_05038: probable nuclear migration protein.
doi:10.1371/journal.pone.0067926.g005

and a low rate of recombination between the two sequenced strains and the gene sequences predicted were also shared with two other Fusarium species, *F. oxysporum* and *F. verticillioides*. Separating these large blocks of conserved DNA, were several smaller regions with high DNA polymorphism, a very high recombination frequency, and these contained many of the predicted gene sequences considered to be unique to *F. graminearum*. These small unique regions of the genome were located in both the sub-telomeric and interstitial regions of each chromosome and were proposed to be the fusion sites of ancestral smaller chromosomes.

Due to the unusual topology of the *F. graminearum* genome landscape, the chromosomal positioning of the predicted virulence genes accross the four *Fusarium graminearum* chromosomes was explored (**Figure 6**). Visual inspection revealed that most of the virulence genes predictions lie in the lower recombination conserved part of the chromosomes (white and blue). However, four predicted virulence genes reside in chromosome regions with a high/very high recombination frequency (4 cM–8 cM, red and >8 cM crimson), namely – FGSG_00071 (**Figure 4**), FGSG_15983, FGSG_04618 (**Figure 5**) and FGSG_16412. Therefore the rarer type of genome landscape is explored in this network analysis. These 4 predicted virulence genes are found in many other species.

Table 5. Comparison of the distribution of known phenotypes of the seeds within the four predicted networks.

Phenotype	Network type			
	Protein-protein interactions	Coexpression	Sequence similarity	Integrated
	Seeds			
Reduced	4	2	4	4
	Predictions			
Essential	5	0	9	11
Reduced	9	3	20	23
Unaffected	11	0	16	22

Counts of the different phenotypes according to the study by Wang and collegues [26] that were found among the predictions derived using four different networks.
doi:10.1371/journal.pone.0067926.t005

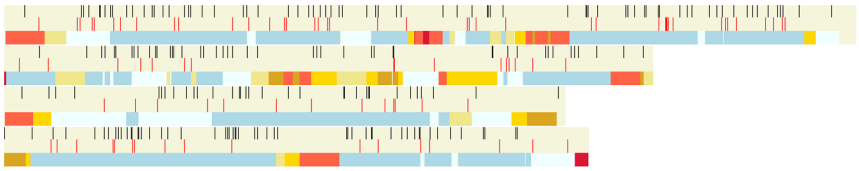


Figure 6. Position of the predictions in relation to the 4 chromosomes of *Fusarium graminearum*. The predicted virulence genes are shown as black vertical bars in track 1 for each chromosome. The verified virulence seeds (red bars) are depicted in track 2. Recombination frequency across the chromosomes is depicted in track 3 using a colour gradient (white (0.0) lowest to crimson (>8 cM highest). The various colours in track 3 for each chromosome indicate the frequence of recombination (cM/27 kb), i. e. # clBeige 1 clKhaki 2 clGold 3 clGoldenRod 4 clTomato 8 clCrimson. The numbers between the colours are boundary values in cM/27 kb. Beige represents the lowest and crimson the highest recombination frequency [47]. (Image generated using OmniMapFree [27]).
doi:10.1371/journal.pone.0067926.g006

The Predicted Virulence Associated Protein Set Shows an Increased Abundance in the Functional Categories Defense/Virulence and Cellular Communication

The functional classification system developed by the Munich Information Centre for Protein Sequences (MIPS) allows the automatic annotation of protein sets into 20 high level functional categories (Funcat) [37] (http://mips.helmholtz-muenchen.de/genre/proj/FGDB/). We hypothesised that successful prediction of virulence associated protein candidates using networks should also increase the annotation frequency of proteins belonging to Funcat groups comprising proteins involved in virulence and protein-protein interactions. Both the protein sets for the seeds and the predicted virulence associated proteins were compared (**Table 6**). A chi-square test showed that both groups are significantly different ($P \geq 0.001$). The Funcat groups 14 (protein fate), 30 (cellular communication/signal transduction mechanism) and 32 (cell rescue, defense and virulence) were increased, while the number of proteins belonging to Funcat group 99 (unclassified proteins and others) was strongly reduced.

Discussion

The integration of multiple types of data such as co-expression, protein-protein interaction and sequence relatedness can provide biological context to particular proteins by showing their relationship to other proteins. In some cases such an approach can provide enhanced annotation or indeed the first annotation for a sequence. For example a protein of unknown function may be strongly co-expressed or may interact with a number of proteins whose functions are known and this may help in narrowing down the possible roles of the previously unannotated protein. Here we used a similar 'guilt by association' approach to examine the network neighbourhood of proteins known to be involved in pathogenicity or virulence for the fungal Ascomycete species *Fusarium graminearum*. There is a large amount of biological, genome and transcriptome information publically available for this species and other pathogenic *Fusarium* species [4,38–40] because of the ever rising economic global importance of Fusarium ear blight disease (www.scabusa.org, [2,41].

This study greatly extends the previous network study of ([12] - Liu et al.). The integrated relationship network developed in this study leads to 215 predictions, of which 29 are hypothetical proteins (as annotated by the Fusarium Database ([6] - Wong et al.) and 25 are fungal specific. The integrated network was particularly informative and predicted 15 proteins linked to virulence that were only revealed in this network. Of these, FGSG_06878 has now been linked to virulence via the shotgun functional analysis of the predicted kinome ([21]-Wang et al.), whilst FGSG_03535 (TRI4) is known to be highly upregulated *in planta* and is required for the synthesis of of the DON mycotoxin. The function of the other 13 predicted virulence associated proteins from the integrated network has not yet been established (1) and/or tested (12). In addition, this study generated four predictions, where the prediction was linked to either 7 or 8 seeds. Of these FGSG_00071 (TRI1), FGSG_07251 and FGSG_10066 have each recently been shown to be required for virulence, whilst the FGSG_09715 single gene deletion mutant was unaffected in pathogenicity towards wheat floral tissue. This level of correct prediction amongst the sequences most highly connected to the verified virulence seeds could be a way of further prioritising the list.

Amongst the 215 predictions, several proteins are considered to have a direct role in virulence because these are required for the production of the DON mycotoxin virulence factor, i.e. example 2. However, the rest of the predictions could play either a direct or indirect role in virulence. The analysis of the sequence type and protein size distribution of the predictions would indicate that this study has underexplored the possible effector component of *Fusarium graminearum*. At the present time we consider most of the predicted virulence associated proteins identified in this study to have an indirect role in virulence and could be seen as system components [18].

One of the caveats with the approach we have taken is that predictions can be biased in favour of nodes with high degree centrality values. The degree centrality of a node in the network is a measure of the number of edges connected to that node, and the higher the value the more 'hub-like' is the node. We used the Kolmogorov-Smirnov test (see, for example, [42]) to compare the (cumulative) distributions of each of the three possible pairs of degree centrality data sets, namely (i) the nodes corresponding to the verified virulence seeds vs. the nodes of the integrated network, (ii) the nodes corresponding to the predicted virulence associated proteins vs. the nodes of the integrated network and (iii) the nodes corresponding to the predicted virulence associated proteins vs. the nodes corresponding to the verified virulence seeds. The test revealed that there was no significant difference for (i) but that there was a highly significant difference for (ii) and (iii). This may reflect a bias in the predictions towards high degree centrality nodes, as such nodes are more likely to be connected to two or more seed proteins.

Another potential limitation of the approach is that for many pathogens (excluding well studied examples such as *Fusarium graminearum* (see for example [7]), *Magnaporthe oryzae*, a rice pathogen and *Ustilago maydis*, a maize pathogen, there is typically very limited information on proteins that have been investigated experimentally for their contribution to virulence and that can act as seeds. Additionally the set of verified virulence seeds is most likely biased with certain types of protein being the subject of more

Table 6. Funcat analysis of the verified virulence seeds and candidate virulence associated proteins.

The main functional categories		Seeds (%)	Candidates (%)
1	metabolism	9.9	7.6
2	energy	0.8	1.6
10	cell cycle and dna processing	4.3	7.1
11	transcription	9.1	4.4
12	protein synthesis	0.4	0.8
14	**protein fate (folding, modification, destination)**	**4.0**	**11.2**
16	protein with binding function or cofactor requirement	10.3	10.3
18	regulation of metabolism and protein function	4.0	4.4
20	cellular transport, transport facilities and transport routes	1.6	4.3
30	**cellular communication/signal transduction mechanism**	**4.0**	**9.3**
32	**cell rescue, defense and virulence**	**1.6**	**4.5**
34	interaction with the environment	3.2	3.7
36	systemic interaction with the environment	0.8	1.2
38	transposable elements, viral and plasmid proteins	0.0	0.1
40	cell fate	4.3	2.6
41	development	2.0	1.2
42	biogenesis of cellular components	3.2	4.4
43	cell type differentiation	6.7	4.2
45	tissue differentiation	0.4	0.3
47	organ differentiation	0.4	0.5
70	Subcellular localization	9.5	9.4
99	**unclassified proteins and others**	**19.8[&]**	**6.9**

[&]recovered from forward genetic screens.

doi:10.1371/journal.pone.0067926.t006

intense biological investigations. For example, for *F. graminearum* although the analysis of the predicted transcription factors and protein kinases (the kinome) has been thoroughly explored ([43] [26] so far the function of the predicted secretome has not [44]. This means that currently the network space is not evenly sampled and may result in many potential targets being missed. Over the next few years this problem could either become worse if the community focusses on genes and gene families already known to essential for infection and/or disease formation in other pathogenic species, or the position may improve as the results from large forward genetic screens for pathogenicity factors and/or via the screening of hypothetical and conserved hypothetical sequences occurs.

Recently, a large scale targeted gene disruption study to search for novel secreted fungal virulence genes was reported for the rice blast pathogen *Magnaporthe orzyae* [45]. In total, 78 putative secreted proteins, most with low sequences similarlity, but highly expressed during the early stages of plant infection, were tested for function. Only one *M. oryzae* gene was shown to be required for virulence in cereal plants. Deletion of the orthologous gene reduced the virulence of another fungal pathogen *Colletotrichum orbiculaire*, which causes anthracnose disease on non-cereal plants. This novel virulence gene has a very restricted fungal taxon distribution. Overall, this recent large experimental biology study reveals just how low a level of predictive success was achieved (1.28%) from an initial highly focussed bioinformatics analyses. Therefore at the present time, the sensitivity of our predictions for *F. graminearum* virulence associated proteins from using the integrated network (1.66%, **Table S4**) is comparable to that achieved using a partially bioinformatically guided, direct experimental approach.

Once genome sequence and gene function information is published on different strains of the same species, several closely related species, or *formae specialis*, then the power of this type of predictive technique is likely to greatly increase. For example, within the Fusaria the number of species under experimental investigation is gradually expanding and involves the use of a range of cereal, non-cereal and mammalian host infecting species. These studies include *F. oxysporum* f.sp *lycopersici* and various other *formae specialis*, which infect different dicotyledonous plant species, *F. solani* as well as *F. verticillioides*, *F. culmorum* and *F. pseudograminearum*, which infect a range of cereal hosts. Also, it is anticipated that in the next five years due to the increased efficiency of generating single gene deletion strains in specific plant pathogenic species, this type of integrated network could be used for comparative analyses involving evolutionarily closely related fungal species with subtly different infection routes and/or host ranges.

The protein interaction component of the integrated network representing predicted interactions [14] was built using known interaction data from 7 non-pathogenic, non-filamentous fungal organisms using information from interologs and domain-domain interactions. Therefore interactions between *Fusarium* specific proteins will not have been captured. The identification within the integrated network of a prediction involved in trichothecene mycotoxin production (**Figure 2**), indicates the value of including co-expresssion data. With the increasing use of next generation sequencing technologies to explore the interaction transcriptome in greater detail, it is conceivable that co-expression information

on different phases of the interaction could be used to further refine the virulence associated protein predictions.

Exploration of the network together with expert biological knowledge about the predicted proteins in the neighbourhoods of verified virulence proteins may lead to a further reduction in hypothesis space and prioritisation to a few genes that could be the target for experimental investigation. However, two separate *F. graminearum* large studies recently published, explored the function of the 709 predicted transcription factors (TAPs) [43] and the 116 predicted protein kinases [26], indicate that the testing of the entire 215 predictions in a focussed project would be feasible via a consortium research approach.

Methods

The Integrated Network

Starting with version 3.3 of PHI-base, *Fusarium graminearum* genes were selected whose contributions to virulence have been tested experimentally and were classified according to whether they have an effect or not. Further expert curation of more recent literature for this study added more *Fusarium graminearum* genes that experiments suggest are involved in virulence and that are currently not in PHI-base Vers. 3.3. **Table S1** shows the complete list of seed genes. In total, these 133 experimentally-tested genes are referred to as the verified virulence (VV) 'seed' genes. The mapping of *Fusarium graminearum* entries in PHI-base to corresponding sequences taken from the latest annotation of the *Fusarium graminearum* genome at the Broad Institute (gene call FG3) was carried out using BLAST and manually reviewed. The total numbers of VV seeds is 100, and the 'virulence unaffected' seeds is 33. The *F. graminearum* genome is predicted to code for 13,332 proteins.

We have described the construction of the integrated network for *Fusarium graminearum* and explored its community structure in [19]. The network was constructed using information from three component data sources, namely gene co-expression, protein sequence similarity and predicted protein-protein interactions. The co-expression component of the network was constructed from the complete publically-available set (12 experiments, 158 individual slides) of *Fusarium* expression studies form PLEXdb [15] that used *Fusarium* Affymetrix GeneChip array [5]. This included 6 *in planta* experiments and 6 *in vitro* studies using the wild-type sequenced PH-1 strain and/or single gene deletion mutants generated in the PH-1 strain on which the GeneChip array was designed (**Table S7**). The data was downloaded in the form of.CEL files, pooled and normalised using the Robust Multichip Average (Irizarry et al., 2003), at which point a data matrix of size 18069 (genes) X 158 (samples) was constructed. The similarity of expression profiles was measured using weighted Pearson correlation coefficient, according to the method of [46]. The sparse network was constructed from the correlation matrix by applying a threshold of 0.88. This value was determined to be optimal for this dataset using the method of Elo et al. (2007), which derives the optimal correlation cut-off value based on the topological properties of the network. The probe set IDs from the FG3 annotation of *Fusarium* [6] were integrated using a mapping file obtained from MIPS (http://mips.helmholtz-muenchen.de/genre/proj/FGDB/). The sequence similarity network was constructed from the results of an all-versus-all sequence matching of the proteins in version 3.2 of the Fusarium annotation at (http://mips.helmholtz-muenchen.de/genre/proj/FGDB/) implemented on a TimeLogic® Tera-BLAST™ (Active Motif Inc., Carlsbad, CA). The network was constructed by creating a "similar sequence" edge joining the two nodes (genes) when there was a pairwise similarity observed between their sequences (bidirectional hit) with expected value of less than 10^{-6}. The co-expression network, the predicted core PPI of Zhao et al [14], the sequence similarity network and the mutant phenotype annotations (from PHI-base and the more recently curated literature) were imported into the Ondex data integration and visualisation system [16] (www.ondex.org) and combined. Merging the nodes that had the same gene accession resulted in the union of the two networks. The coexpression values and scores derived from BLAST were included as weights on appropriate edges and are included in the final integrated network available with this paper. The explanation about how the BLAST scores were calculated and the distribution of these values for all edges used in predictions are included as a **Figure S6**. It is, therefore possible to adjust the threshold further in Ondex network visualisation software and explore what effects it would have on the network and the predictions.

In this study we were interested in the potential of the network for prediction. The prediction of virulence genes was achieved by implementing a new plug-in software module for the Ondex system. The plug-in works by creating a set of sub-graphs that include genes annotated to be of relevance to virulence (the verified virulence seeds) and their nearest neighbours with respect to co-expression, PPI and sequence similarity in the constituent and combined networks. The genes were predicted to be likely important for virulence if there were at least two known virulence-relevant genes found in their immediate network neighbourhood, in a similar manner to that of Liu et al [12]. The seed nodes, the predictions and the edges connecting predictions to seeds were "tagged" to create gene lists, which could then be used to select relevant subsets of the network for visualisation in the graphical user interface of Ondex.

The Ondex software can be downloaded from www.ondex.org. The integrated network, seed genes and predictions are made available in **File S1**.

Supporting Information

Figure S1 The entire integrated network containing the predicted virulence associated gene FGSG_06878 connected to 8 verified virulence seeds.
(DOCX)

Figure S2 The local integrated network containing the predicted virulence associated gene 07251 connected to 7 verified virulence seeds.
(DOCX)

Figure S3 The integrated network containing the predicted virulence associated gene FGSG_09715 connected to 7 verified virulence seeds.
(DOCX)

Figure S4 The integrated network containing the predicted virulence associated gene FGSG_10066 connected to 7 verified virulence seeds.
(DOCX)

Figure S5 The integrated network containing the predicted virulence associated gene FGSG_00472 connected to 5 verified virulence seeds.
(DOCX)

Figure S6 The distribution of e-values for sequence similarity edges that were used for deriving predictions.
(DOCX)

Table S1 List of 133 seed verified virulence (VV) genes.
(DOCX)

Table S2 Selected annotation for the 215 predicted virulence associated proteins.
(XLSX)

Table S3 The ratios of seed associated to all other edges for all of the proteins predicted to be associated with virulence.
(DOCX)

Table S4 Estimating the predictive power of the four different networks.
(DOCX)

Table S5 Heatmap showing the taxonomic diversity of the matches to the 215 predictions.
(DOCX)

Table S6 Prediction of FGSG_09715, FGSG_07251 and FGSG_10066 as virulence associated proteins.
(DOCX)

Table S7 Mapping the data from Wang et al to the integrated network.
(DOCX)

Table S8 The publically available *F. graminearum* microarray gene expression datasets used in this study.
(DOCX)

File S1 ZIP archive file for Ondex containing the integrated network, seed genes and predictions in OXL format.
(ZIP)

Author Contributions

Conceived and designed the experiments: MS KHK AL MU ST EJS. Analyzed the data: AL KHK MU MS LB. Wrote the paper: KHK MS AL MU LB CR.

References

1. Goswami RS, Kistler HC (2004) Heading for disaster: Fusarium graminearum on cereal crops. Mol Plant Pathol 5: 515–525.
2. Dean R, Van Kan JA, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, et al. (2012) The Top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol 13: 414–430.
3. Yuen GY, Schoneweis SD (2007) Strategies for managing Fusarium head blight and deoxynivalenol accumulation in wheat. Int J Food Microbiol 119: 126–130.
4. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, et al. (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science 317: 1400–1402.
5. Guldener U, Mannhaupt G, Munsterkotter M, Haase D, Oesterheld M, et al. (2006) FGDB: a comprehensive fungal genome resource on the plant pathogen *Fusarium graminearum*. Nucleic Acids Res 34: D456–458.
6. Wong P, Walter M, Lee W, Mannhaupt G, Munsterkotter M, et al. (2011) FGDB: revisiting the genome annotation of the plant pathogen Fusarium graminearum. Nucleic Acids Res 39: D637–639.
7. Urban M, Hammond-Kosack KE (2012) Molecular genetics and genomic approaches to explore Fusarium infection on wheat floral tissue. In: Brown D, Proctor RH, editors. Fusarium: genomics and molecular and cellular biology. Norwich: Horizon Scientific Press. 43–79.
8. Winnenburg R, Urban M, Beacham A, Baldwin TK, Holland S, et al. (2008) PHI-base update: additions to the pathogen host interaction database. Nucleic Acids Res 36: D572–576.
9. Winnenburg R, Baldwin TK, Urban M, Rawlings C, Kohler J, et al. (2006) PHI-base: a new database for pathogen host interactions. Nucleic Acids Res 34: D459–464.
10. Baldwin TK, Winnenburg R, Urban M, Rawlings C, Koehler J, et al. (2006) The pathogen-host interactions database (PHI-base) provides insights into generic and novel themes of pathogenicity. Mol Plant Microbe Interact 19: 1451–1462.
11. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306: 1555–1558.
12. Liu X, Tang WH, Zhao XM, Chen L (2010) A network approach to predict pathogenic genes for *Fusarium graminearum*. PLoS One 5: e13021.
13. Wolfe CJ, Kohane IS, Butte AJ (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics 6: 227.
14. Zhao XM, Zhang XW, Tang WH, Chen L (2009) FPPI: *Fusarium graminearum* protein-protein interaction database. J Proteome Res 8: 4714–4721.
15. Wise RP, Caldo RA, Hong L, Shen L, Cannon E, et al. (2007) BarleyBase/PLEXdb. Methods Mol Biol 406: 347–363.
16. Kohler J, Baumbach J, Taubert J, Specht M, Skusa A, et al. (2006) Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 22: 1383–1390.
17. Lysenko A, Hindle MM, Taubert J, Saqi M, Rawlings CJ (2009) Data integration for plant genomics–exemplars from the integration of *Arabidopsis thaliana* databases. Briefings in Bioinformatics 10: 676–693.
18. Schneider DJ, Collmer A (2010) Studying Plant-Pathogen Interactions in the Genomics Era: Beyond Molecular Koch's Postulates to Systems Biology. In: VanAlfen NKBGLJE, editor. Annual Review of Phytopathology, Vol 48. 457–479.
19. Bennett L, Lysenko A, Papageorgiou L, Urban M, Hammond-Kosack K, et al. (2012) Detection of multi-clustered genes and community structure for the plant pathogenic fungus *Fusarium graminearum*. Lect Notes Comp Science 7605: 69–86.
20. Hagberg AA SD, Swart PJ. (2008) Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux, G., Vaught, T., Millman, J. (eds.) Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA USA, 11–15.
21. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. Fast unfolding of communities in large networks Journal of Statistical Mechanics: Theory and Experiment 2008: 100008.
22. Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69: 026113.
23. Chen J, Yuan B (2006) Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics 22: 2283–2290.
24. Liu G, Wong L, Chua HN (2009) Complex discovery from weighted PPI networks. Bioinformatics 25: 1891–1897.
25. Rispail N, Soanes DM, Ant C, Czajkowski R, Grunler A, et al. (2009) Comparative genomics of MAP kinase and calcium-calcineurin signalling components in plant and human pathogenic fungi. Fungal Genetics and Biology 46: 287–298.
26. Wang C, Zhang S, Hou R, Zhao Z, Zheng Q, et al. (2011) Functional analysis of the kinome of the wheat scab fungus Fusarium graminearum. PLoS Pathogens 7: e1002460.
27. Antoniw J, Beacham AM, Baldwin TK, Urban M, Rudd JJ, et al. (2011) OmniMapFree: a unified tool to visualise and explore sequenced genomes. BMC Bioinformatics 12: 447.
28. Beacham A (2011) Pathogenicity of *Fusarium graminearum* and *Fusarium culmorum* on wheat ears. PhD thesis in Plant Pathogen and Microbiology (Exeter University, Exeter).
29. Seong KY, Pasquali M, Zhou X, Song J, Hilburn K, et al. (2009) Global gene regulation by *Fusarium* transcription factors Tri6 and Tri10 reveals adaptations for toxin biosynthesis. Mol Microbiol 72: 354–367.
30. Dyer RB, Plattner RD, Kendra DF, Brown DW (2005) *Fusarium graminearum TRI14* is required for high virulence and DON production on wheat but not for DON synthesis in vitro. J Agricult Food Chem 53: 9281–9287.
31. McCormick SP, Alexander NJ, Harris LJ (2010) *CLM1* of *Fusarium graminearum* encodes a longiborneol synthase required for culmorin production. Appl Environ Microbiol 76: 136–141.
32. Gardiner DM, Kazan K, Manners JM (2009) Nutrient profiling reveals potent inducers of trichothecene biosynthesis in *Fusarium graminearum*. Fungal Genet Biol: 604–613.
33. Desjardins AE (2006) *Fusarium* Mycotoxins - Chemistry, Genetics and Biology. St. Paul, Minnesota U.S.A.: The American Phytopathological Society.
34. McCormick SP, Harris LJ, Alexander NJ, Ouellet T, Saparno A, et al. (2004) Tri1 in Fusarium graminearum encodes a P450 oxygenase. Appl Environ Microbiol 70: 2044–2051.
35. Brown NA, Bass C, Baldwin TK, Chen H, Massot F, et al. (2011) Characterisation of the *Fusarium graminearum*-wheat floral interaction. J Pathogens 2011.
36. Yu HY, Seo JA, Kim JE, Han KH, Shim WB, et al. (2008) Functional analyses of heterotrimeric G protein G alpha and G beta subunits in *Gibberella zeae*. Microbiology 154: 392–401.
37. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, et al. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32: 5539–5545.
38. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, et al. (2009) The genome of Nectria haematococca: contribution of supernumerary chromosomes to gene expansion. PLoS Genetics 5: e1000618.
39. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, et al. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464: 367–373.
40. Rep M, Kistler HC (2010) The genomic organization of plant pathogenicity in Fusarium species. Curr Opin Plant Biol 13: 420–426.

41. Nunes CC, Dean RA (2011) Host-induced gene silencing: a tool for understanding fungal host interaction and for developing novel disease control strategies. Mol Plant Pathol 13: 519–529.
42. Siegel S (1956) Nonparametric Statistics for the behavioural Sciences: McGraw-Hill, New York, US.
43. Son H, Seo YS, Min K, Park AR, Lee J, et al. (2011) A phenome-based functional analysis of transcription factors in the cereal head blight fungus, *Fusarium graminearum*. PLoS Pathog 7: e1002310.
44. Brown NA, Antoniw J, Hammond-Kosack KE (2012) The predicted secretome of the plant pathogenic fungus Fusarium graminearum: a refined comparative analysis. PLoS One 7: e33731.
45. Saitoh H, Fujisawa S, Mitsuoka C, Ito A, Hirabuchi A, et al. (2012) Large-Scale Gene Disruption in Magnaporthe oryzae Identifies MC69, a Secreted Protein Required for Infection by Monocot and Dicot Fungal Pathogens. PLoS Pathog 8: e1002711.
46. Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, et al. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Res 35: D863–869.
47. Gale LR, Bryant JD, Calvo S, Giese H, Katan T, et al. (2005) Chromosome complement of the fungal plant pathogen *Fusarium graminearum* based on genetic and physical mapping and cytological observations. Genetics 171: 985–1001.